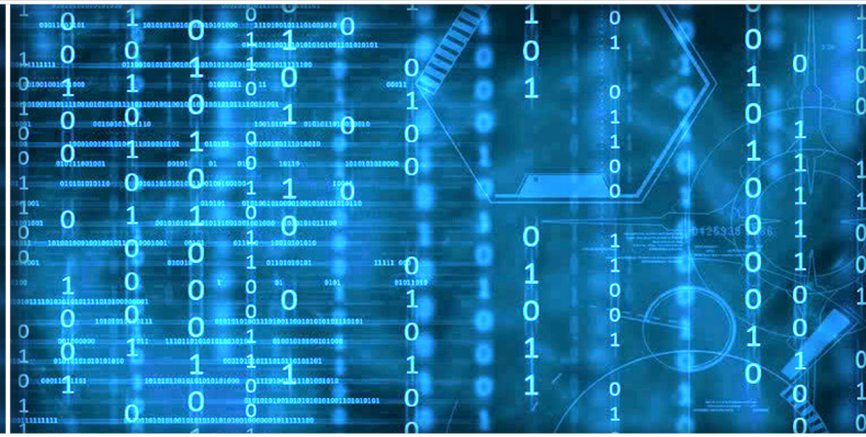


Volume 12 Issue 4

April 2021



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 12 Issue 4 April 2021
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: Road Detection Method based on Online Learning

Authors: Wenbo Wang, Yong Ma

PAGE 1 – 7

Paper 2: Autonomous Reusing Policy Selection using Spreading Activation Model in Deep Reinforcement Learning

Authors: Yusaku Takakuwa, Hitoshi Kono, Hiromitsu Fujii, Wen Wen, Tsuyoshi Suzuki

PAGE 8 – 15

Paper 3: The Adoption of Mobile Health Applications by Patients in Developing Countries: A Systematic Review

Authors: Nasser Aljohani, Daniel Chandran

PAGE 16 – 21

Paper 4: Method for Most Appropriate Plucking Date Determination based on the Elapsed Days after Sprouting with NIR Reflection from Sentinel-2 Data

Authors: Kohei Arai, Yoshiko Hokazono

PAGE 22 – 29

Paper 5: Systems Security Affection with the Implementation of Quantum Computing

Authors: Norberto Novoa Torres, Juan Carlos Suarez Garcia, Erik Alexis Valderrama Guancha

PAGE 30 – 39

Paper 6: Energy Storage and Electric Vehicles: Technology, Operation, Challenges, and Cost-Benefit Analysis

Authors: Surender Reddy Salkuti

PAGE 40 – 45

Paper 7: Detecting Unauthorized Network Intrusion based on Network Traffic using Behavior Analysis Techniques

Authors: Nguyen Tung Lam

PAGE 46 – 51

Paper 8: An Integrated Implementation Framework for an Efficient Transformation to Online Education

Authors: Ahmed Al-Hunaiyyan, Salah Al-Sharhan, Rana Alhajri, Andrew Bima

PAGE 52 – 61

Paper 9: Development and Print of Clothing through Digitalized Designs of Natural Patterns with Flexible Filaments in 3D Printers

Authors: Jean Roger Farfán Gavancho, Wilber Antonio Figueroa Quispe, Dayvis Victor Farfán Gavancho, Beto Puma Huamán, Victor Manuel Lima Condori, George Jhonatan Cahuana Alca

PAGE 62 – 76

Paper 10: Smartphone-based Recognition of Human Activities using Shallow Machine Learning

Authors: Maha Mohammed Alhumayyani, Mahmoud Mounir, Rasha Ismael

PAGE 77 – 85

Paper 11: Deep Learning Approaches for Intrusion Detection in IIoT Networks – Opportunities and Future Directions

Authors: Thavavel Vaiyapuri, Zohra Sbai, Haya Alaskar, Nourah Ali Alaseem

PAGE 86 – 92

Paper 12: Sensing and Detection of Traffic Status through V2V Routing Hop Count and Route Energy

Authors: Mahmoud Zaki Iskandarani

PAGE 93 – 100

Paper 13: Annotated Corpus of Mesopotamian-Iraqi Dialect for Sentiment Analysis in Social Media

Authors: Al-Khafaji Ali J Askar, Nilam Nur Amir Sjarif

PAGE 101 – 105

Paper 14: Exploring Parkinson's Disease Predictors based on Basic Intelligence Quotient and Executive Intelligence Quotient

Authors: Haewon Byeon

PAGE 106 – 111

Paper 15: Is Deep Learning Better than Machine Learning to Predict Benign Laryngeal Disorders?

Authors: Haewon Byeon

PAGE 112 – 117

Paper 16: Distance Education during COVID-19 Pandemic: The Perceptions and Preference of University Students in Malaysia Towards Online Learning

Authors: Husna Hafiza Razami, Roslina Ibrahim

PAGE 118 – 126

Paper 17: VG4 Cipher: Digital Image Encryption Standard

Authors: Akhil Kaushik, Vikas Thada

PAGE 127 – 133

Paper 18: Formulation of Association Rule Mining (ARM) for an Effective Cyber Attack Attribution in Cyber Threat Intelligence (CTI)

Authors: Md Sahrom Abu, Siti Rahayu Selamat, Robiah Yusof, Aswami Ariffin

PAGE 134 – 143

Paper 19: Integrated Pairwise Testing based Genetic Algorithm for Test Optimization

Authors: Baswaraju Swathi, Harshvardhan Tiwari

PAGE 144 – 150

Paper 20: Iterative Decoding of Chase Pyndiah Decoder Utilizing Multiple Relays Network

Authors: Saif E. A. Alnawayseh

PAGE 151 – 156

Paper 21: An Efficient Privacy Preserving Approach for e-Health

Authors: Supriya Menon M, Rajarajeswari Pothuraju

PAGE 157 – 162

Paper 22: Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features

Authors: Oscar Utomo Kumala, Amalia Zahra

PAGE 163 – 168

Paper 23: NetAI-Gym: Customized Environment for Network to Evaluate Agent Algorithm using Reinforcement Learning in Open-AI Gym Platform

Authors: Varshini Vidyadhar, Nagaraj R, D V Ashoka

PAGE 169 – 176

Paper 24: Service Outages Prediction through Logs and Tickets Analysis

Authors: Sunita A Yadwad, V. Valli Kumari, S Venkata Lakshmi

PAGE 177 – 183

Paper 25: Recent Themes of Colombian Scientific Engineering Journals in Scopus

Authors: Marco Aguilera-Prado, Octavio José Salcedo Parra, Eduardo Avendaño Fernández

PAGE 184 – 189

Paper 26: Book Recommendation for Library Automation Use in School Libraries by Multi Features of Support Vector Machine

Authors: Kitti Puritat, Phichete Julrode, Pakinee Ariya, Sumalee Sangamuang, Kannikar Intawong

PAGE 190 – 196

Paper 27: Econometric Analysis of Stock Market Performance during COVID-19 Pandemic: A Case Study of Uzbekistan Stock Market

Authors: Mansur Eshov, Walid Osamy, Ahmed Aziz, Ahmed M. Khedr

PAGE 197 – 204

Paper 28: Deep Learning based Anomaly Detection in Images: Insights, Challenges and Recommendations

Authors: Ahad Alloqmani, Yoosef B. Abushark, Asif Irshad Khan, Fawaz Alsolami

PAGE 205 – 215

Paper 29: The Evaluation of User Experience Testing for Retrieval-based Model and Deep Learning Conversational Agent

Authors: Pui Huang Leong, Ong Sing Goh, Yogan Jaya Kumar, Yet Huat Sam, Cheng Weng Fong

PAGE 216 – 221

Paper 30: Survey of Tools and Techniques for Sentiment Analysis of Social Networking Data

Authors: Sangeeta Rani, Nasib Singh Gill, Preeti Gulia

PAGE 222 – 232

Paper 31: Efficient Security Model for RDF Files Used in IoT Applications

Authors: Mohamed El kholy, Abdel baes Mohamed

PAGE 233 – 239

Paper 32: Predictive Analysis of Ransomware Attacks using Context-aware AI in IoT Systems

Authors: Vyitarani Mathane, P.V. Lakshmi

PAGE 240 – 244

Paper 33: Contribution to the Improvement of Cryptographic Protection Methods for Medical Images in DICOM Format through a Combination of Encryption Method

Authors: Maka Maka Ebenezer, Pauné Félix, Malong Yannick, Simo Ntso Pascal Junior, Nnemé Nnemé Léandre

PAGE 245 – 250

Paper 34: Birds Identification System using Deep Learning

Authors: Suleyman A. Al-Showarah, Sohyb T. Al-qbailat

PAGE 251 – 260

Paper 35: Feature Engineering Algorithms for Traffic Dataset

Authors: Akibu Mahmoud Abdullah, Raja Sher Afgun Usmani, Thulasyammal Ramiah Pillai, Ibrahim Abaker Targio Hashem, Mohsen Marjani

PAGE 261 – 268

Paper 36: PlexNet: An Ensemble of Deep Neural Networks for Biometric Template Protection

Authors: Ashutosh Singh, Ranjeet Srivastva, Yogendra Narain Singh

PAGE 269 – 280

Paper 37: A Multi-layer Machine Learning-based Intrusion Detection System for Wireless Sensor Networks

Authors: Nada M. Alruhaily, Dina M. Ibrahim

PAGE 281 – 288

Paper 38: ParaDist-HMM: A Parallel Distributed Implementation of Hidden Markov Model for Big Data Analytics using Spark

Authors: Imad Sassi, Samir Anter, Abdelkrim Bekkhoucha

PAGE 289 – 303

Paper 39: Performance Assessment of Context-aware Online Learning for Task Offloading in Vehicular Edge Computing Systems

Authors: Mutaz A. B. Al-Tarawneh, Saif E. Alnawayseh

PAGE 304 – 320

Paper 40: Body Weight Estimation using 2D Body Image

Authors: Rohan Soneja, Prashanth S, R Aarthi

PAGE 321 – 326

Paper 41: A Detailed Study on the Choice of Hyperparameters for Transfer Learning in Covid-19 Image Datasets using Bayesian Optimization

Authors: Miguel Miranda, Kid Valeriano, Jos´e Sulla-Torres

PAGE 327 – 335

Paper 42: Impact of Deep Learning on Localizing and Recognizing Handwritten Text in Lecture Videos

Authors: Lakshmi Haritha Medida, Kasarapu Ramani

PAGE 336 – 344

Paper 43: A Vehicle Routing Problem for the Collection of Medical Samples at Home: Case Study of Morocco

Authors: Effazi Haitam, Rafalia Najat, Jaafar Abouchabaka

PAGE 345 – 351

Paper 44: A Computer-Assisted Collaborative Reading Model to Improve Reading Fluency of EFL Learners in Continuous Learning Programs in Saudi Universities

Authors: Abdulfattah Omar, Mohamed Saad Mahmoud Hussein, Fahd Shehail Alalwi

PAGE 352 – 359

Paper 45: Optimal Allocation of DG and D-STATCOM in a Distribution System using Evolutionary based Bat Algorithm

Authors: Surender Reddy Salkuti

PAGE 360 – 365

Paper 46: Analysis of Load Variation Consideration for Optimal Distributed Generation Placement

Authors: Aida Fazliana Abdul Kadir, Mohamad Fani Sulaima, Noor Ropidah Bujal, Mohd Nazri Bin Abd Halim, Elia Erwani Hassan

PAGE 366 – 372

Paper 47: Resource Utilization Prediction in Cloud Computing using Hybrid Model

Authors: Anupama K C, Shivakumar B R, Nagaraja R

PAGE 373 – 381

Paper 48: An Experiment for Outdoor GPS Localization Enhancement using Kalman Filter with Multiantenna Consumer-Grade Sensors

Authors: Phudinan Singkhamfu, Parinya Suwansrikham

PAGE 382 – 388

Paper 49: Artificial Intelligence Model based on Grey Clustering for Integral Analysis of Industrial Hygiene Risk

Authors: Alexi Delgado, Diana Aliaga, Cristian Carlos, Lisseth Vergaray, Chiara Carbajal

PAGE 389 – 395

Paper 50: Intelligent Traffic Light Controller using Fuzzy Logic and Image Processing

Authors: Abdelkader Chabchoub, Ali Hamouda, Saleh Al-Ahmadi, Adnen Cherif

PAGE 396 – 399

Paper 51: Private LTE Network Service Management Model, based on Agile Methodologies, for Big Mining Companies

Authors: José Valdivia-Bedregal, Norka Bedregal-Alpaca, Elisa Castañeda-Huaman

PAGE 400 – 406

Paper 52: Speeding up Natural Language Text Search using Compression

Authors: Majed AbuSafiya

PAGE 407 – 409

Paper 53: Developing an IoT Platform for the Elderly Health Care

Authors: Medhat Awadalla, Firdous Kausar, Razzaqul Ahshan

PAGE 410 – 417

Paper 54: Study and Analysis for the Choice of Optical Fiber in the Implementation of High-Capacity Backbones in Data Transmission

Authors: Wilmer Vergaray-Mendez, Brian Meneses-Claudio, Alexi Delgado

PAGE 418 – 428

Paper 55: On State-of-the-art of POS Tagger, 'Sandhi' Splitter, 'Alankaar' Finder and 'Samaas' Finder for Indo-Aryan and Dravidian Languages

Authors: Hema Gaikwad, Jatinderkumar R. Saini

PAGE 429 – 436

Paper 56: The Development of Students' Spatial Orientation through the use of 3D Graphics

Authors: Benjamín Maraza-Quispe

PAGE 437 – 443

Paper 57: Symptoms-Based Fuzzy-Logic Approach for COVID-19 Diagnosis

Authors: Maad Shatnawi, Anas Shatnawi, Zakarea AlShara, Ghaith Husari

PAGE 444 – 452

Paper 58: An Optimization Approach for Multiple Sequence Alignment using Divide-Conquer and Genetic Algorithm

Authors: Arunima Mishra, Sudhir Singh Soam, Bipin Kumar Tripathi

PAGE 453 – 458

Paper 59: Proposed Design of White Sugar Industrial Supply Chain System based on Blockchain Technology

Authors: Ratna Ekawati, Yandra Arkeman, Suprihatin, Titi Candra Sunarti

PAGE 459 – 465

Paper 60: Fraud Detection in Shipping Industry using K-NN Algorithm

Authors: Ganesan Subramaniam, Moamin A. Mahmoud

PAGE 466 – 475

Paper 61: Generating Test Cases using Eclipse Environment – A Case Study of Mobile Application

Authors: Rosziati Ibrahim, Nurul Ain Aswini Abdul Jan, Sapiee Jamel, Jahari Abdul Wahab

PAGE 476 – 483

Paper 62: Traffic Accidents Detection using Geographic Information Systems (GIS)

Authors: Wesam Alkhadour, Jamal Zraqou, Adnan Al-Helali, Sajeda Al-Ghananeem

PAGE 484 – 494

Paper 63: An Agent-Based Evaluation Model of Students' Emotional Engagement in Classroom

Authors: Moamin A. Mahmoud, Latha Subramanian, Ihab L Hussein Alsammak, Mahmood H. Hussein

PAGE 495 – 505

Paper 64: Wireless Body Area Sensor Networks for Wearable Health Monitoring: Technology Trends and Future Research Opportunities

Authors: Malek ALRASHIDI, Nejah NASRI

PAGE 506 – 512

Paper 65: A-SA SOS: A Mobile- and IoT-based Pre-hospital Emergency Service for the Elderly and Village Health Volunteers

Authors: Kannikar Intawong, Waraporn Boonchieng, Peerasak Lertrakarnnon, Ekkarat Boonchieng, Kitti Puritat

PAGE 513 – 518

Paper 66: An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks

Authors: P. Bharath Kumar Chowdary, R. Udaya Kumar

PAGE 519 – 525

Paper 67: Identification of Babbitt Damage and Excessive Clearance in Journal Bearings through an Intelligent Recognition Approach

Authors: Joel Pino Gómez, Fidel E. Hernández Montero, Julio C. Gómez Mancilla, Yenny Villuendas Rey

PAGE 526 – 533

Paper 68: An Empirical Investigation of the Relationship Between Business Process Transparency and Business Process Attack

Authors: Alhanouf Aldayel, Ahmad Alturki

PAGE 534 – 545

Paper 69: A Tree-profile Shape Ultra Wide Band Antenna for Chipless RFID Tags

Authors: A K M Zakir Hossain, Nurulhalim Bin Hassim, Jamil Abedalrahim Jamil Alsayaydeh, Mohammad Kamrul Hasan, Md. Rafiqul Islam

PAGE 546 – 550

Paper 70: Non-Hodgkin Type Lymphoma Cancer Cell Detection using Connected Components Labeling and Moments of Image

Authors: Monirul Islam Pavel, Mohsinul Bari Shakir, Dewan Ahmed Muhtasim, Omar Faruk

PAGE 551 – 556

Paper 71: Grey Clustering Method for Water Quality Assessment to Determine the Impact of Mining Company, Peru

Authors: Alexi Delgado, Jhoel Andy Gauna Achata, Jorge Alfredo Barreda Valdivia, Julio Cesar Junior Santivañez Montes, Chiara Carbajal

PAGE 557 – 564

Paper 72: Analysis of Speech Signal Data of Missing Vowels using Logistic Regression and K-Means Clustering

Authors: Ujjal Saikia, Jiten Hazarika

PAGE 565 – 570

Paper 73: Hybrid SFLA-UBS Algorithm for Optimal Resource Provisioning with Cost Management in Multi-cloud Computing

Authors: Muhammad Iftikhar Hussain, JingSha He, Nafei Zhu, Fahad Sabah, Zulfiqar Ali Zardari, Saqib Hussain, Fahad Razque

PAGE 571 – 578

Paper 74: Software Engineering Ethics Competency Gap in Undergraduate Computing Qualifications within South African Universities of Technology

Authors: Senyeki M. Marebane, Robert T. Hans

PAGE 579 – 592

Paper 75: Healthcare Logistics Optimization Framework for Efficient Supply Chain Management in Niger Delta Region of Nigeria

Authors: Imeh J. Umoren, Ubong E. Etuk, Anietie P. Ekong, Kingsley C. Udonyah

PAGE 593 – 604

Paper 76: A Data Science Framework for Data Quality Assessment and Inconsistency Detection

Authors: Anusuya Ramasamy, Berhanu Sisay, Amanuel Bahiru

PAGE 605 – 613

Paper 77: Investigation of Smart Home Security and Privacy: Consumer Perception in Saudi Arabia

Authors: Omar Almutairi, Khalid Almarhabi

PAGE 614 – 622

Paper 78: Smart Company System using Hybrid Nomenclature of Neural Network

Authors: Mbida Mohamed, Ezzati Abdellah

PAGE 623 – 631

Paper 79: Handling Sudden and Recurrent Changes in Business Process Variability: Change Mining based Approach

Authors: Asmae HMAMI, Hanae SBAI, Mounia FREDJ

PAGE 632 – 640

Paper 80: The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem

Authors: Mustafa Abdul Salam, Ahmad Taher Azar, Mustafa Samy Elgendy, Khaled Mohamed Fouad

PAGE 641 – 655

Paper 81: A Sentiment Analysis of Egypt's New Real Estate Registration Law on Facebook

Authors: Abdulfattah Omar, Wafya Ibrahim Hamouda

PAGE 656 – 663

Paper 82: Factors Affecting Mobile Learning Acceptance in Higher Education: An Empirical Study

Authors: Nahil Abdallah, Odeh Abdallah, OM Bohra

PAGE 664 – 671

Paper 83: Novel Properties for Total Strong - Weak Domination Over Bipolar Intuitionistic Fuzzy Graphs

Authors: As'ad Mahmoud As'ad Alnaser

PAGE 672 – 679

Paper 84: Performance Comparison of Three Hybridization Categories to Solve Multi-Objective Flow Shop Scheduling Problem

Authors: Jebari Hakim, Siham Rekiek, Rahali El Azzouzi Saida, Samadi Hassan

PAGE 680 – 689

Paper 85: Learners Classification for Personalized Learning Experience in e-Learning Systems

Authors: A. JOHN MARTIN, M. MARIA DOMINIC, F. Sagayaraj Francis

PAGE 690 – 697

Paper 86: Efficient Security Solutions for IoT Devices

Authors: Faleh Alfaleh, Salim Elkhediri

PAGE 698 – 707

Paper 87: Internet of Things (IoT) based Smart Vehicle Security and Safety System

Authors: Yassine SABRI, Aouad Siham, Aberrahim Maizate

PAGE 708 – 714

Paper 88: Permissioned Blockchain: Securing Industrial IoT Environments

Authors: Samira Yeasmin, Adeel Baig

PAGE 715 – 725

Paper 89: A Hybrid Technique based on RSA and Data Hiding for Securing Handwritten Signature

Authors: Yaser Maher Wazery, Shima Gamal Haridy, AbdElmegeid Amin Ali

PAGE 726 – 735

Paper 90: Design and Performance Measurement of Energy-based Acoustic Signal Detection with Autonomous Underwater Vehicles

Authors: Redouane Es-sadaoui, Jamal Khallaayoune, Tamara Brizard

PAGE 736 – 745

Paper 91: A New Corner Detection Operator for Multi-Spectral Images

Authors: Hassan El Houari, Ahmed Fouad El Ouafdi

PAGE 746 – 751

Paper 92: Fast Fractal Coding of MRI Images using Deep Reinforcement Learning

Authors: Bejoy Varghese, S. Krishnakumar

PAGE 752 – 759

Paper 93: Comprehensive Analysis of Two Malicious Arabic-Language Twitter Campaigns

Authors: Reem Alharthi, Areej Alhothali, Kawthar Moria

PAGE 760 – 771

Paper 94: Integrating Cost-231 Multiwall Propagation and Adaptive Data Rate Method for Access Point Placement Recommendation

Authors: Fransiska Sisilia Mukti, Puput Dani Prasetyo Adi, Dwi Arman Prasetya, Volvo Sihombing, Nicodemus Rahanra, Kristia Yuliawan, Julianto Simatupang

PAGE 772 – 777

Paper 95: A Comparative Analysis of Hadoop and Spark Frameworks using Word Count Algorithm

Authors: Yassine Benlachmi, ABDELAZIZ EL YAZIDI, MOULAY LAHCEN HASNAOUI

PAGE 778 – 788

Paper 96: Security Aspects of Electronic Health Records and Possible Solutions

Authors: Prashant Vilas Kanade, Arun Kumar

PAGE 789 – 793

Road Detection Method based on Online Learning

Wenbo Wang¹

Nanjing Audit University
NanJing, China 211899

Yong Ma²

Nanjing University of Science and Technology
NanJing, China 210014

Abstract—Road detection is always the key problem of researches on areas of unmanned ground vehicle and computer vision. A road detection method is proposed based on online learning and multi-sensor fusion. First of all, the Lidar point clouds are projected onto the images via the joint calibration of these two kinds of sensors. Then Simple Linear Iterative Clustering is used to segment images into many superpixels. Based on that, a multilayer online learning method is proposed, in which 2 Support Vector Machines are trained to detect the road. To be specific, the superpixel layer Support Vector Machine is used to detect road roughly, and the pixel layer Support Vector Machine is then trained to classify the edge pixels of the road areas, which is classified by the upper-layer Support Vector Machine. These 2 Support Vector Machines are updated online at each frame to be adapted to the changing environment. At last, some experiments are carried out on KITTI RAW dataset and an autonomous land vehicle, and the results show the effectiveness of proposed method. The main contributions of this work lie on as follows: 1) a multilayer learning model is proposed to detect road more robustly and accurately; 2) an online learning method is proposed which can be adapted to the changing environment.

Keywords—Road detection; data fusion; unmanned ground vehicle; online learning; image segmentation

I. INTRODUCTION

Road detection [1], [2], [3] is one of the key technologies in multiple research areas such as unmanned ground vehicle development and machine vision. Traditional road detection methods are based on image data captured by RGB cameras mostly [4]. In the past decades, researchers have provided bunches of image processing algorithms [5], [6], [7], such as vanishing point localization, natural road boundaries detection or CRF [8] (conditional random fields) optimization [9] after segmentation of pixels based on prior information. However, different illumination conditions, shadow and complicated texture background significantly impact on image quality, which leads to rapid decline of algorithm performance. Recently, with the development of 3 dimensional sensor, researchers have put forward various road detection methods based on range data. For example, Sunando Sengupta and Paul Sturgess constructed an octree model to describe the detected environment and attached an advanced CRF model for semantic segmentation [10]. Benjamin Suger published a semi-supervised machine learning method, using Lidar to construct an accessible probability map, for outdoor navigation of robots [11].

Compared to RGB cameras, three dimensional sensors outperforms in many areas. Range data of surrounding is detected from all directions based on 3D sensors [12], [13], providing adequate information of target structure without interference from illumination conditions, severe shadow, complicated texture background and etc. However, 3D sensors have

their own weaknesses. First, taking stereoscopic vision into account such as Kinect, they are influenced easily by moving targets, leading to large amount of noise in detected range image [14], [15]. Furthermore, their observation range is more than 20 meters usually, which could not fulfil the requirements of unmanned ground vehicle environment perception. Second, although the observation range of Lidar is 120 meter in maximum, the detecting data is increasingly sparse with the increasing distance, due to its fixed angular resolution on measuring targets. So, Lidar sensor could not describe complete terrain or target detail appearance in distance [16], [17].

In order to combine the advantages of these two kinds of sensors, increasing number of researchers fuse the outcomes of these two to acquire better environment model, especially on road detection [18], [19]. In addition, due to the continuous surrounding changes of unmanned ground vehicle, it is difficult for the pre-trained classifier to perform well on road detection and classification. In order to deal with this problem, unmanned ground vehicles are required to do online learning [20], [21] based on real-time surround environment. In this paper, a hierarchical online learning method is put forward and verified its efficiency on KITTI raw dataset and our own unmanned ground vehicle. The main innovation of this paper is as follows. (1) A hierarchical model is introduced, which could realize robust road detection more accurate. (2) An online learning method is put forward, which is adaptive to continuous environment changes.

The main content of this paper is organized as follows: In Section II, we briefly review the fusion of image and lidar data. In Section III, we will propose our model in detail. In Section IV, several experiments are designed to verify the effectiveness of our model. In Section V, some conclusions are given to finish this paper.

II. FUSION OF IMAGE AND LIDAR DATA

To combine image and range data, joint calibration is needed for data alignment of camera and Lidar. Lidar point cloud is expressed as $\mathbf{P}_{lidar} = \{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ and its corresponding projection outcome is recorded as $\mathbf{P}_{image} = \{U, V\}$

$$\begin{aligned} \mathbf{P}_{image} &= \mathbf{R}_{rect}^0 \mathbf{T}_{lidar}^{image} \mathbf{P}_{lidar} \\ \mathbf{T}_{lidar}^{image} &= \begin{pmatrix} \mathbf{R}_{lidar}^{image} & \mathbf{t}_{lidar}^{image} \\ 0 & 1 \end{pmatrix} \end{aligned} \quad (1)$$

In which, \mathbf{R}_{rect}^0 R is the matrix to transform original image visual angle to front view, while image lidar $\mathbf{T}_{lidar}^{image}$ is the matrix to project Lidar point to image view. $\mathbf{R}_{lidar}^{image}$ and $\mathbf{t}_{lidar}^{image}$ are rotation matrix and translation vector to project Lidar point

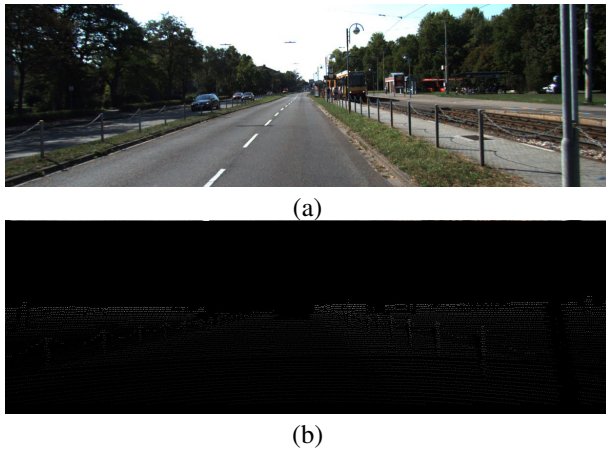


Fig. 1. Result of Lidar Point Clouds Projection. Subfigure (a) is the Original Image Captured by RGB Camera, while Subfigure (b) is the Lidar Projection Outcome.

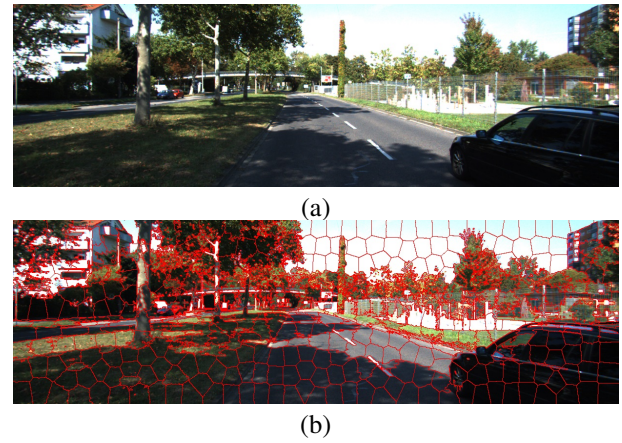


Fig. 3. Result of Super-pixel Segmentation based on SLIC.

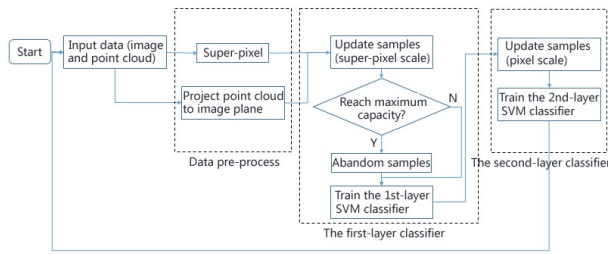


Fig. 2. Multiple Layers Online Learning Model.

cloud to image layer respectively. In many accessible datasets, such as KITTI dataset, calibration parameters are in open access, so we skip the detail computation procedure of these parameters. Details of joint calibration computation could be found in reference [22].

After joint calibration mentioned above, Lidar point cloud is projected to image plane, as Fig. 1 shown. Subfigure (a) is the original image captured by RGB camera, while subfigure (b) is the Lidar projection outcome, in which height feature is indicated by brightness. With joint calibration and projection procedures, some pixels on image is corresponding to Lidar point, which means their range and height is available.

III. MULTIPLE LAYERS ONLINE LEARNING MODEL

A. Hierarchical Online Learning Definition

In order to adapt to continuous changing background environment in road detection, we put forward a multiple layers online learning model as Fig. 2 shown.

In the rest of this paper, we will introduce this model in details.

B. Super-pixel and Road Detection Definition

Super-pixel is a series of adjacent pixels composed of small areas with similar color, brightness and texture characteristics. Most of these small areas retain the effective information for further image segmentation, and generally do not destroy the

boundary information of objects in the image. Therefore, more and more image segmentation algorithms adopt super-pixel as the basic segmentation unit [21], [23]. As Fig. 3 shown, SLIC super-pixel method is utilized to several super-pixel units in this paper. In super-pixel procedure, segmentation is expressed as $R = \{r_i, i=1..M\}$, in which M is the number of super-pixels. Then, road detection can be considered as a bi-classification problem:

$$\begin{aligned} r_i &= (F_i, L_i) \\ F_i &= \{f_{i1}, f_{i2}, \dots, f_{id}\} \\ L_i &= \{+1, -1\} \end{aligned} \quad (2)$$

Here, F_i is the corresponding characteristic of r_i , d is the feature dimension and L_i represents the final classification outcome, in which $L_i=+1$ indicates that rF_i belongs to road region, and $L_i=-1$ means that r_i is an off-road region.

C. Choice of Classifier

In recent years, deep learning and other new machine learning methods have been extensively studied, and have achieved remarkable results on many datasets. However, deep learning requires a large number of training samples to learn network parameters, and requires very high computational resources. Even with the use of transfer learning and other learning technologies, the deep learning method is difficult to apply in the specific task scenario of online road detection. Compared with deep learning method, the final decision function of Support Vector Machine (SVM)[24] classifier is determined only by a few support vectors rather than all training samples. Therefore, the computational complexity is low and key samples can be calculated automatically to realize efficient machine learning. In the real-life scene using online road learning, the sample size is often small, and the distribution of positive and negative samples is very uneven in the initial stage of learning. Therefore, compared with other classifiers, SVM has excellent performance in online learning system. In addition, in view of the problem that the computational complexity of SVM classifier will raise with the increase of training samples, the online learning model proposed in this paper sets a maximum sample size and an update strategy



Fig. 4. Road Detection Result Only via Lidar Data.

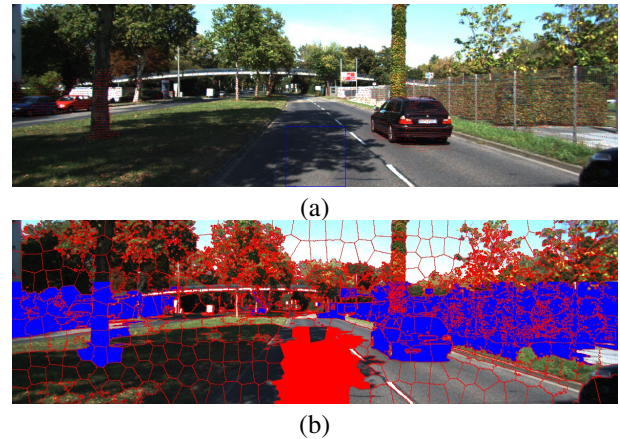


Fig. 5. Selection of Training Samples of Classifier in the First Layer.

of positive and negative samples to ensure that the proposed method can meet the real-time requirements of road detection.

D. Road Detection in Super-pixel Level

In the multi-level learning model proposed in this paper, the first level classifier is at super-pixel scale. Generally speaking, one of the important factors affecting the performance of online road detection methods is the selection strategy of positive and negative samples. Many image-based online road detection algorithms assume a small area at the middle-bottom of the image as an initial road area and the image edge belonging to a non-road area [9], [21] when selecting classifier samples. However, in the actual scene, the edge of the image is mostly sky or building, which can not represent all kinds of negative samples near road. On the contrary, in many road detection methods based on Lidar, there are some problems in extracting positive samples. A typical algorithm for road detection based on Lidar data is to project the point cloud data of Lidar into two-dimensional grid map, then calculate the height difference of each grid. By setting a threshold artificially, the grid whose height difference is lower than the threshold is taken as a positive sample and the other grid as a negative sample [25]. As shown in Fig. 4, the first picture is the original image captured by RGB camera, and the area labelled in red in the second image indicates the positive sample area. In such methods, off-road regions with small height difference (such as lawn and road) can not be effectively distinguished.

In conclusion, image data is more suitable for extracting positive samples, while Lidar data is more suitable for extracting negative ones. Therefore, in this paper, we use a fusion method in this paper to combine Lidar and image data. Specifically, we assume that the super-pixels in the rectangular frame (as shown in Fig. 5(a)) at the image bottom are positive samples. Then use the Lidar data to separate obstacles from non-obstacles, and assume that the super-pixels belonging to obstacles are negative samples.

In order to figure out the super-pixels which belong to obstacles, Lidar point cloud data is projected to a plane constructed by X and Y coordinate axis. The plane is described in grid map form, then the maximum height difference of each grid is calculated and an artificial threshold is set to separate road grids and obstacle grids. Afterwards, project all obstacle grids to image plane as shown in red pixels in Fig.

5(a). Finally, those super-pixels with over a threshold of Lidar obstacle pixels are labelled as negative samples. As can be seen in Fig. 5(b) positive and negative samples are colored in red and blue respectively.

For each super-pixel unit, the color histogram and LBP texture features in HSI space are extracted from the color image, and the average height feature and height variance feature are extracted from the projected Lidar point image. After that, these samples are added to the sample library, and the first-level SVM classifier is updated and trained online. Then, all the super-pixel units in the whole image are classified by this classifier, and the first-level road detection results are obtained.

In addition, the online learning model sets the total capacity of the training sample library when updating the first level classifier online for each frame. If the current training sample library is not full, the positive and negative samples of this frame are directly added to the training sample library. If the current training sample library capacity has reached the maximum, then according to the proportion of positive and negative samples in the current sample library and sample collection time, part of the samples are deleted to make room for samples selected from current frame. Specifically, if there are more positive samples than negative samples in the current sample bank, the positive samples will be deleted, and on the other hand, the negative samples will be deleted. Secondly, the oldest samples will be deleted first to ensure that the trained classifier can continuously adapt to the latest environment.

E. Road Boundary Classification in Pixel Scale

In the super-pixel-based road detection algorithm, different scale setting of super-pixels has a significant impact on the accuracy of road detection. Many researchers are trying to solve this problem by using the idea of stratification. For example, document [21] proposes a multi-scale learning framework. They set up θ (odd) super-pixel segmentation layers of different scales from small to large, then trained a SVM classifier at each level, and used a voting method to get the final classification results next. However, in our research procedure, scale of super-pixels only influence road boundary pixels, while inner road regions are not sensitive to super-pixel scale as can be

seen in Fig. 6. In each image, super-pixel segmentation is shown in upper half, while road detection outcome in the below half. In that, this kind of methods witness a narrow promotion after consuming large amount of computation resources. To reduce the complexity of algorithm as well as ensure the accuracy of boundary localization, we train a second layer of SVM classifier in pixel scale based on road boundary super-pixels to polish up road detection result, after acquiring the first layer segmentation procedure. To be specific, we suppose road super-pixel set as R_{road} , which is segmented by the first layer classifier, while E_i indicates whether i th super-pixel belongs to road boundary, and its neighbourhood is expressed as η_i . Then E_i is calculated as follow.

$$E_i = \begin{cases} 1, & \text{if } \exists r_j \notin R_{road} \& r_j \in \eta_i \\ 0, & \text{else} \end{cases} \quad (3)$$

In the second SVM classifier layer, all pixels in road boundary super-pixels ($E_i = 1$) are become samples to be classified. Here, all pixels on road ($E_i = 0$) are used as positive samples, while pixels in obstacle super-pixels in the first layer are used as negative samples.

Unlike the updating method of the first-layer classifier, the second-layer classifier does not retain the samples of historical frames, but extracts the RGB values of all positive and negative samples of the current frame, as well as the average height and height variance of the super-pixels as features from each frame. We retrain the classifier in this layer, and segments all samples to be classified. Fig. 7 shows an example of sample choosing for the second-layer classifier. In Fig. 7, red and blue region indicates all the positive and negative sample pixels, respectively to train the second-layer SVM classifier, while the green region indicates all the pixels to be classified (road boundary region).

Finally, after the first-layer super-pixel classifying, the second-layer of our method separate road and off-road pixels in road boundary regions (green areas in Fig. 7). Combining output of the first-layer, the final road detection result is observed. The efficiency of the edge polishing up procedure is proved by experiments afterwards.

IV. EXPERIMENTS

To verify the efficiency of proposed algorithm, we choose a dataset (2011_09_26_drive_0013) randomly from KITTI RAW DATA. This dataset contains 143 continuous frames with a resolution of 1242×375 pixels, as well as corresponding Lidar frame with over 100 thousand points and the joint calibration parameters. The KITTI RAW DATA does not provide ground truth, so we label road regions on each frame artificially for algorithm verification.

Fig. 8 shows the number of training samples of the first-layer classifier with time accumulating. As can be seen in Fig. 8, positive samples are less than negative ones at the very beginning in the super-pixel classifier layer with a training bank containing 5000 training samples in total. However, with time going, the difference between positive and negative sample number starts to decrease after the 47th frame. The oldest sample keeps to be replaced continuously, and reaches

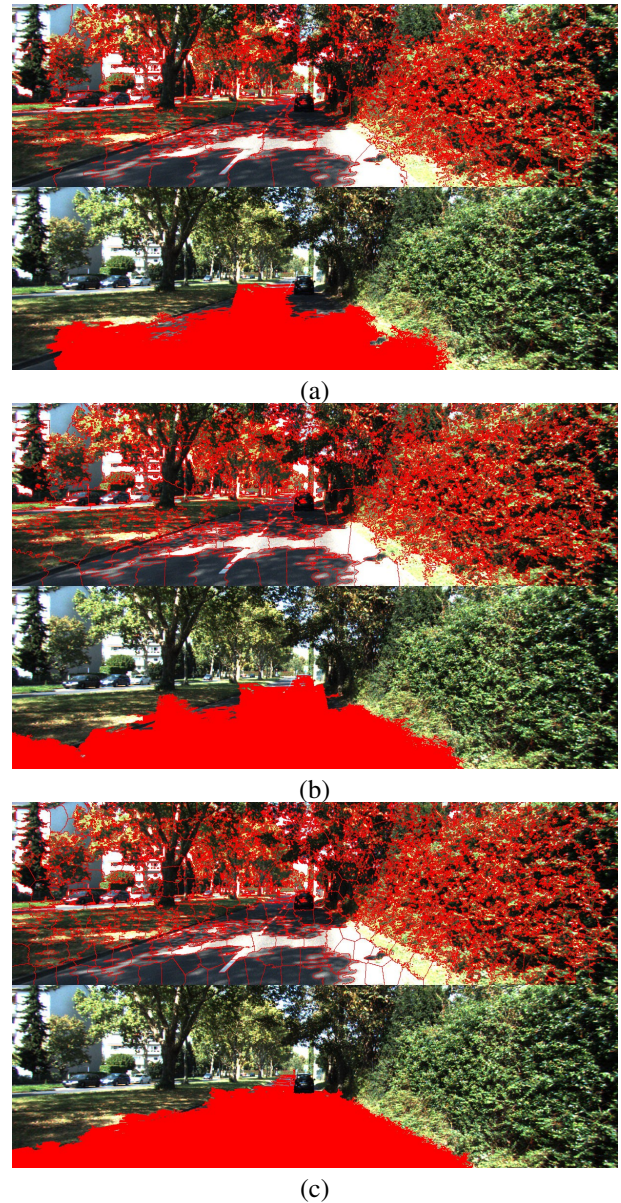


Fig. 6. Road Detection Results in Super-pixel Layer of Different Scales. Super-pixel Number is Set at 100 in (a), 200 in (b) and 500 in (c).

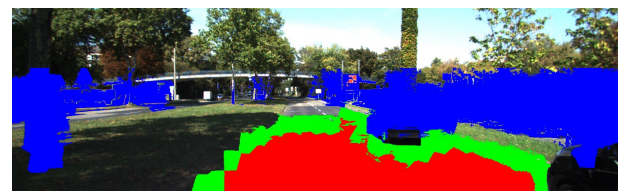


Fig. 7. Selection of Samples of Classifier in Pixel Layer.

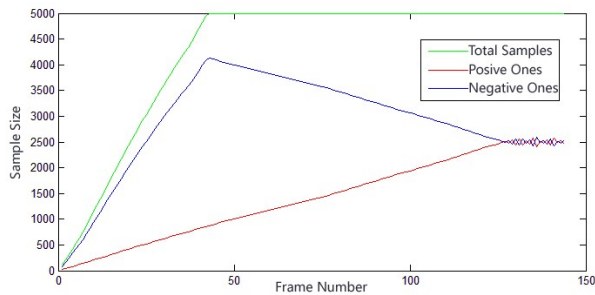


Fig. 8. Number of Training Samples of the SVM Classifier in the First Layer.

a balance with almost the same percentage. An uneven classifying problem becomes an even one at this point.

Fig. 9 illustrates road detection results of single layer (first layer) and multiple layers (total structure) learning models. As can be seen in Fig. 9, multiple layers model outperforms the single layer model. Due to an independent classification procedure of road boundary pixels, it can make full use of the road boundary information and detect road border well, which lead to higher road detection accuracy.

In order to further verify the efficiency of multi-sensors fusion and hierarchical online learning model, we realize four kinds of road detection methods on KITTI RAW DATA:

(1) Compute height difference in each super-pixel using Lidar data [25], then set a threshold (25 cm) artificially. Label super-pixels with height difference lower than this threshold as road regions, while the rest as non-road regions.

(2) Attach the method proposed in document [21] and use multi-scale super-pixel voting model for road detection.

(3) Use the fusion method proposed in this paper, by extracting positive and negative sample. Then detect road combining the method proposed in document [21], which uses multi-scale super-pixel voting model to detect road.

(4) Use the proposed multi-sensors fusion method in this paper to extract positive and negative samples, then detect road by multi-layers online learning model.

In this paper, we verify these four kinds of online road detection models' efficiency by six different parameters: FPR, TPR, Precision, Recall, Accuracy and F-measure. Table I put forward all kinds of parameters of these four learning models.

As can be seen in Table I, model (3) witness a significant promotion on road detection efficiency compared to model (1) and (2), so that the multi-sensors fusion efficiency can be proved compared to the single sensor road detection. In addition, Model (4) achieves the best performance, which further proves that multi-layer online learning model makes efforts to better road detection.

Next, we test the proposed method on our own unmanned ground vehicle to verify the efficiency in actual driving scenario. The unmanned ground vehicle is shown in Fig. 10. This vehicle contains a pre-calibrated RGB camera, Velodyne HDL64 Lidar and other kinds of sensors.

In our experiments, 3000 frames are collected in total as a dataset. Each frame contains a RGB image and corresponding

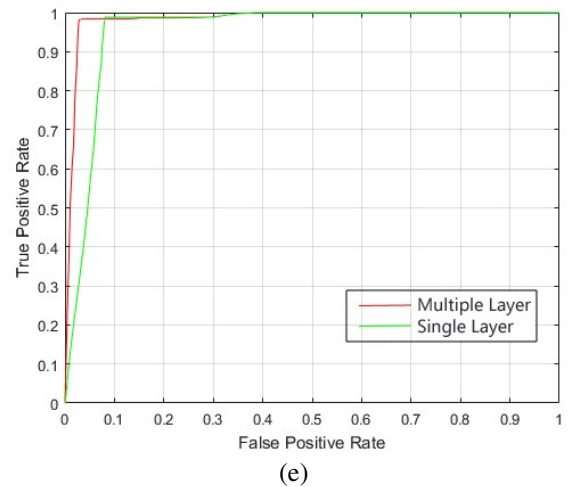
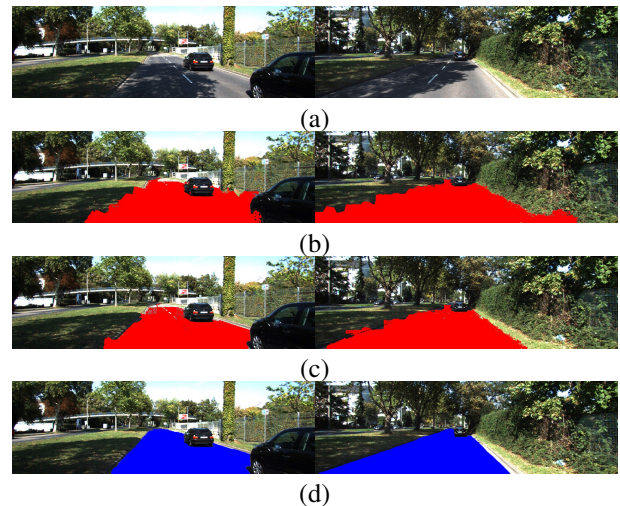


Fig. 9. Road Detection Results in Super-pixel Layer of Different Scales. (a) illustrates the Original Image of Two Input Scenario, (b) Shows the Outcome of the First-layer, (c) Indicates the Outcome of the Second-layer, and (d) is the Ground Truth, while (e) is the ROC Curve of Single Layer and Multiple Layers Models.

TABLE I. PERFORMANCES OF DIFFERENT ROAD DETECTION METHODS

Model	FPR	TPR	Precision	Recall	Accuracy	F-measure
Model (1)	0.0846	0.8043	0.6517	0.8043	0.8161	0.7593
Model (2)	0.0740	0.9301	0.7033	0.9301	0.9241	0.7941
Model (3)	0.0734	0.9313	0.7126	0.9313	0.9276	0.8033
Model (4)	0.0680	0.9466	0.7178	0.9466	0.9343	0.8165

Lidar point cloud data, as well as ground truth with artificial labelled road region. Finally, the proposed method achieves 91.07% precision on this dataset with an average running time at 87.35ms per frame, which meets real-time requirements. Fig. 11 shows part of our road detection results.

V. CONCLUSION

This paper proposes a road detection model used on unmanned ground vehicle based on online learning and multi-sensors fusion. According to our model, SLIC method is first utilized to separate image data to several super-pixels. Then Lidar data which belongs to obstacles is projected to image



Fig. 10. An Autonomous Land Vehicle.

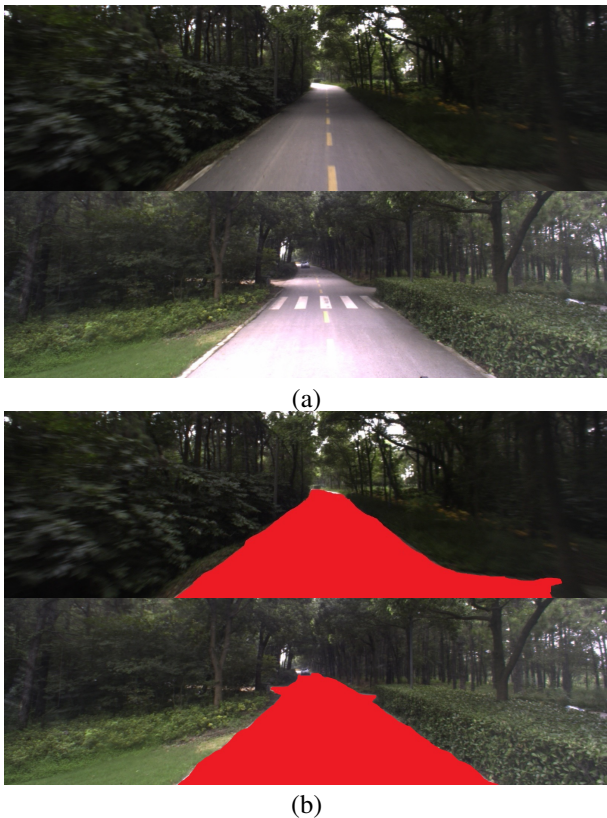


Fig. 11. Some Experimental Results. (a) Proposes Two Input Images, while
(b) Indicates the Corresponding Road Detection Outcomes.

plane to separate image super-pixels into two kinds: obstacles and non-obstacles. Here in our method, we assume that super-pixels at mid-bottom of the image belongs to road region.

Afterwards, we put forward a multi-layer online learning model. In the first layer, large scale road detection is fulfilled by a SVM classifier, which trained by road and obstacle super-pixels. Next, another SVM classifier is developed for meticulous road detection in boundary regions. We utilize a new strategy to update the training sample bank, which could balance the percentage of positive and negative samples auto-

matically. Maximum sample amount is limited to deal with the distribution problem of training data. Real-time requirements are met, while hierarchical classifier online learning is also accomplished to adapt to the environment changes.

The experiments performed on KITTI RAW DATA and our unmanned ground vehicle confirm that the proposed method meets real-time requirements in online learning road detection.

REFERENCES

- [1] B. S. Shin, Z. Xu, and R. Klette, "Visual lane analysis and higher-order tasks: a concise review," *Machine Vision & Applications*, vol. 25, no. 6, pp. 1519–1547, 2014.
- [2] V. Castillo, A. Díaz-Lvarez, M. Callejo, and F. Serradilla, "Grammar guided genetic programming for network architecture search and road detection on aerial orthophotography," *Applied Sciences*, vol. 2020, no. 10, 2020.
- [3] T. Almeida, B. Loureno, and V. Santos, "Road detection based on simultaneous deep learning approaches," *Robotics and Autonomous Systems*, vol. 133, p. 103605, 2020.
- [4] W. U. Hua-Yue and L. R. Duan, "Unstructured road detection method based on rgb entropy and improved region growing," *Journal of Jilin University(Engineering and Technology Edition)*, 2019.
- [5] H. Kong, J. Y. Audibert, and J. Ponce, "General road detection from a single image," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 19, no. 8, p. 2211, 2010.
- [6] J. M. Alvarez, A. M. Lopez, T. Gevers, and F. Lumberreras, "Combining priors, appearance, and context for road detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1168–1178, 2014.
- [7] C. F. Ndez, R. Izquierdo, D. F. Llorca, and M. A. Sotelo, "Road curb and lanes detection for autonomous driving on urban scenarios," in *IEEE International Conference on Intelligent Transportation Systems*, 2014.
- [8] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*, 2001.
- [9] K. Lu, L. Jian, X. An, and H. He, "A hierarchical approach for road detection," in *IEEE International Conference on Robotics & Automation*, 2014.
- [10] S. Sengupta and P. Sturgess, "Semantic octree: Unifying recognition, reconstruction and representation via an octree constrained higher order mrf," in *IEEE ICRA*, 2015.
- [11] B. Suger, B. Steder, and W. Burgard, "Traversability analysis for mobile robots in outdoor environments: A semi-supervised learning approach based on 3d-lidar data," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015, pp. 3941–3946, 2015.
- [12] K. Rao and J. He, "Analysis on defect characteristics of urban road detection based on 3d ground penetrating radar," *Urbanism and Architecture*, 2019.
- [13] S. Gu, T. Lu, Y. Zhang, A. Jose, J. Yang, and H. Kong, "3d lidar + monocular camera: an inverse-depth induced fusion framework for urban road detection," *IEEE Transactions on Intelligent Vehicles*, vol. PP, no. 3, pp. 1–1, 2018.
- [14] J. Smisek, M. Jancosek, and T. Pajdla, "3d with kinect," *Advances in Computer Vision & Pattern Recognition*, vol. 21, no. 5, pp. 1154–1160, 2011.
- [15] T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of noise in kinect depth images: A review," *Sensors Journal IEEE*, vol. 14, no. 6, pp. 1731–1740, 2014.
- [16] Altmann, Maccarone, McCarthy, Aongus, Newstadt, Gregory, Buller, S. Gerald, and McLaughlin, "Robust spectral unmixing of sparse multispectral lidar waveforms using gamma markov random fields," *IEEE Transactions on Computational Imaging*, vol. PP, no. 99, pp. 1–1, 2016.
- [17] Z. Xiao, H. Li, D. Zhou, Y. Dai, and B. Dai, "Accurate extrinsic calibration between monocular camera and sparse 3d lidar points without markers," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, 2017.

- [18] J. Sock, J. Kim, J. Min, and K. Kwak, "Probabilistic traversability map generation using 3d-lidar and camera," in *IEEE International Conference on Robotics & Automation*, 2016.
- [19] L. Xiao, B. Dai, L. D. T. Hu, and T. Wu, "Crf based road detection with multi-sensor fusion," in *Intelligent Vehicles Symposium*, 2015, pp. 192–198.
- [20] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain," in *Robotics: Science & Systems II, August, University of Pennsylvania, Philadelphia, Pennsylvania, Usa*, 2006.
- [21] Q. Zhang, L. Yong, Y. Liao, and W. Yue, "Traversable region detection with a learning framework," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015, pp. 1678–1683, 2015.
- [22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.
- [23] H. Wang, G. Yan, L. Yong, and M. Ren, "Road detection via superpixels and interactive image segmentation," in *IEEE International Conference on Cyber Technology in Automation*, 2014.
- [24] El-Naqa, Issam, Yongyi, Yang, Wernick, Miles, N., Galatsanos, Nikolas, and P., "A support vector machine approach for detection of microcalcifications." *IEEE Transactions on Medical Imaging*, 2002.
- [25] M. Suzuki, E. Terada, T. Saitoh, and Y. Kuroda, "Vision based far-range perception and traversability analysis using predictive probability of terrain classification," in *ISR/ROBOTIK 2010, Proceedings for the joint conference of ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics)*, 7-9 June 2010, Munich, Germany - Parallel to AUTOMATICA, 2011.

Autonomous Reusing Policy Selection using Spreading Activation Model in Deep Reinforcement Learning

Yusaku Takakuwa¹

Department of Information and Communication Engineering
Tokyo Denki University
Tokyo, Japan

Hitoshi Kono²

Department of Engineering
Tokyo Polytechnic University
Kanagawa, Japan

Hiromitsu Fujii³

Department of Advanced Robotics
Chiba Institute of Technology
Chiba, Japan

Wen Wen⁴

Department of Precision Engineering
The University of Tokyo
Tokyo, Japan

Tsuyoshi Suzuki⁵

Department of Information and Communication Engineering
Tokyo Denki University
Tokyo, Japan

Abstract—This paper describes a policy transfer method of a reinforcement learning agent based on the spreading activation model of cognitive psychology. This method has a prospect of increasing the possibility of policy reuse, adapting to multiple tasks, and assessing agent mechanism differences. In the existing methods, policies are evaluated and manually selected depending on the target–task. The proposed method generates a policy network that calculates the relevance between policies in order to select and transfer a specific policy that is presumed to be effective based on the current situation of the agent while learning. Using a policy network graph structure, the proposed method decides the most effective policy while repeating probabilistic selection, activation, and spread processing. In the experiment section, this study describes experiments conducted to evaluate usefulness, conditions of use, and the usable range of the proposed method. Tests using CartPole and MountainCar, which are classical reinforcement learning tasks, are described and transfer learning is compared between the proposed method and a Deep Q–Network without transfer. As the experimental results, usefulness was suggested in the transfer learning of the same task without manual compared with previous method with various conditions.

Keywords—Reinforcement learning; transfer learning; deep learning; cognitive psychology; spreading activation theory

I. INTRODUCTION

In recent years, practical realization of robots, which can perform flexibly in various environments like those of a human being, is being expected. It is difficult for robots to act flexibly in unknown and dynamic environments without human control. In addition, it is also difficult to establish a control strategy beforehand. Therefore, many researches that allow the robot to learn by itself are being actively conducted [1] [2]. In these researches, reinforcement learning is used for robot

autonomous learning. Reinforcement learning is a method that allows an agent (hereinafter, a learning robot or system will be referred to as an agent) to learn the optimal action by trial and error [3]. Reinforcement learning enables agents to autonomously acquire behavior rules, however there are still some problems that need to be addressed, such as the extensive learning time required for practical and complex tasks. In order to solve this problem, a learning method of reusing a policy of a previously learned task (source–task) for a new task to be learned (target–task) is proposed. This learning method, called transfer learning [4], has been studied in various domains. This method enables agents to improve the adaptability of the target–tasks, thus shortening the learning time. However, most of the transfer learning in the existing research needs to previously determine the policy of the source–task that is considered to be effective for the target–task. Therefore, it is necessary to manually evaluate the policies.

In the existing methods, selecting and reusing effective policies from a plurality of them has been proposed [5] [6]. However, it is considered difficult adapting to obstacles when reusing policies. Specifically, many obstacles should be considered, such as performing different tasks (heterogeneous tasks), utilizing different agent mechanisms (heterogeneous agents), and setting the parameters of reinforcement learning itself, among others. Assuming that agents are used in various fields and environments, it is very difficult for people to set up reinforcement learning parameters considering all obstacles. Therefore, this research discusses a new policy reuse method with a prospect to reduce the manual labor required for policy selection and that can perform flexibly in the presence of various obstacles.

For the above problems, in order to flexibly select a policy,

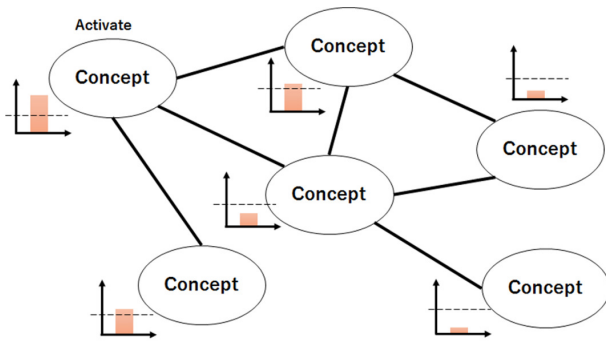


Fig. 1. Simplified Image of the Spreading Activation Model. A Value Called Activation Value Diffuses Through the Network Extending from the Activated Concept.

we apply the spreading activation model as a psychology model that aids human behavior and judgment in the proposed policy selection method based on previous method [7]. In addition, Deep Q-Network method is adopted as function approximation of policy of reinforcement learning, and proposed method is evaluated with various classical reinforcement learning tasks in computer simulation.

A. Spreading Activation and Existing Research

The spreading activation model [8] is a psychology model related to human concept formation (remembrance, re-recognition, etc.) on the assumption that concepts acquired by humans are stored as a network structure in the brain. Many concepts are memorized using schematic representation as network structure in human brain. The concept has an activation value that can be activated or deactivated by external stimuli such as visual information. In the process, activation value increases beyond the threshold value, the concept remain. This phenomena is called recall. Activated values spread to related concepts that are connected via path of semantic distance. In the spreading activation model, there is a concept called semantic distance in which the distance between concepts varies according to the strength of the relevance between these concepts. Concept activation is done via a relevance network. An example of an activation diffusion model is shown in Fig. 1.

Kono *et al.* are proposed transfer reinforcement learning with spreading activation model which is called SAP-net to select the policies adaptively according to environments [7]. SAP-net is discussed effectiveness by simplified computer simulation such as shortest path problem, and it is defined theoretically for implementation. However SAP-net is not consider the computer resources for actual implementation in the robot or agent, which means that it is adopted table type Q function to be described a policy.

B. Policy Description Method

When reusing policies acquired through reinforcement learning, it is important to describe and store them. In reinforcement learning, policy description methods can be roughly divided into two types. One is to describe the behavioral value obtained by learning in a Q function as Q-table which is constructed by look-up table. It is indirectly prescripts the

policy. In the case of using the Q-table, learning is possible by describing all the state-behavior pairs of the agent. By mapping the behavior value from the Q-table for each state, it is possible to perform the transfer learning. However, when using a Q-table, the table size increases exponentially as the number of states increases, which is disadvantageous. Therefore, it is difficult to learn tasks handling many state numbers like in real environments and complicated tasks.

As a second policy description method, there is a method to approximate the behavior value by a function. By function approximation, it is possible to learn tasks with a larger number of states than when using Q-tables. Various methods such as tile coding [9] [10], fuzzy [11] [12], RBFN [13] [14], and deep neural networks [15] have been proposed as function approximation methods. In this research, we discuss a function approximation method using neural networks (NN), which have been actively researched recently, such as Deep Q-Networks, Deep Reinforcement Learning [16] [17] [18] and so on. By function approximation, it is possible to learn tasks with more states than when using a Q-table. However, depending on the setting of the intermediate layer of a NN and the activation function when performing function approximation, it may cause an increase in learning time, excessive learning, and unlearning. In the case of a task performed by an agent in a real environment, it has been considered that transfer learning using a function approximation method is effective; however, it is not realistic to optimize the network structure of a NN for each learning environment, scale, and task.

C. The Principal Aim of this Study

From the above context, the principal aim of this study is to discuss the following two topics.

- 1) The proposed method of automatically selecting effective policies from the policy features, such as the setting of the hidden layers of a NN.
- 2) The prospects of the proposed method, conditions of use, and applicable range.

As a function approximation with reinforcement learning method, Deep Q-Network is implemented in this study. Therefore proposed method is based on policy selection method in transfer reinforcement learning with spreading activation model proposed by Kono *et al.* [7], and is tuned that the policy selection method can be implemented by Deep Q-Network as a part of reinforcement learning algorithm. In the experiment, CartPole and MountainCar are adopted as evaluation function, which are classical reinforcement learning tasks.

II. METHODS

A. Precondition

In this section, to introduce the proposed method, we describe related terms and assumptions. For the reinforcement learning algorithm, Q learning was used.

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha\{r + \gamma \max_{a'} Q(s', a')\} \quad (1)$$

Here, observable state is $s \in S$, and action of agent is $a \in A$. S and A are assumed discrete state. Learning parameters

are learning rate α , discount rate γ and reward r . In addition, hereinafter, the NN model used for function approximation is called policy.

B. Flow of the Proposed Method

Initially, the proposed method classifies each policy into categories based on features of multiple policies learned in advance, and creates a network using these categories. The proposed method selects a policy from the selection probability calculated based on a parameter, which is called the activation value, given to each policy. Policies are selected by referring to selection probabilities. Based on the loss between the value calculated for each action obtained through the policy and the value that acts as the teacher, the action obtained by the selected policy is classified as one that promotes learning (positive transition) or one that does not promote learning (negative transition). From this judgment result, the network is constructed based on the selected policies. By performing this process for each action, the activation value of each policy is changed during the transition learning. Thus, a system that assigns preferential learning to policies with large activation values is constructed.

C. Categorize Policies

Based on the features of the policies, multiple policies are categorized. A category in this research refers to a set of relevant policies that are calculated from the relevance of multiple policies. Policies are evaluated from one viewpoint and judged whether they belong to the same category. This viewpoint is called a prototype. In this study, the prototype is determined empirically based on the information available from the features included in the learned policies. For each category, a policy distance d_{ij} , which describes the relationship between the policies, is generated. The inter-policy distance combines all the policies to cover all the connection patterns in the category.

Multiple benefits can be obtained by classifying policies into categories. First, it is possible to summarize and manage similar policies using a NN of multiple learned policies. It is thought that policies having similar structures of NN layer number and unit number are likely to obtain similar learning results when the other parameters are unified.

Second, we can classify policies learned by with heterogeneous tasks or heterogeneous agents. Policies learned by different tasks and agent mechanisms are difficult to learn with different learning settings. Therefore, categories are useful for organizing and managing policies.

D. Policy Network (SAP-Net)

A policy network is created using the classified categories. The generated policy network is called SAP-Net in this research. SAP-Net is an undirected graph, and it is defined as follows.

$$\mathbb{G} = (\Pi, E, \omega) \quad (2)$$

where Π is the set of policies to be combined, E shows the path connection relation between policies, and ω is the

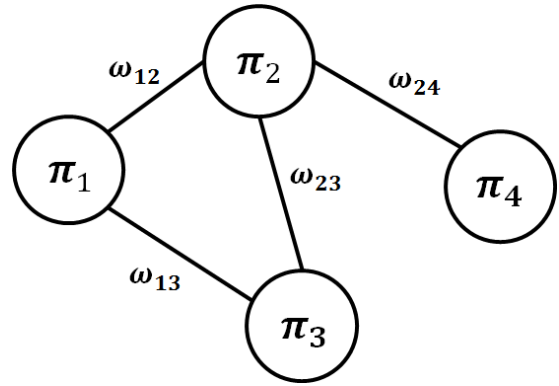


Fig. 2. SAP-Net. Multiple Policies are Tied by Policy Distance d_{ij} .

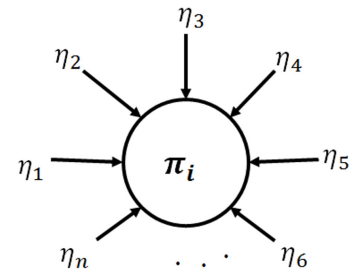


Fig. 3. Activation Value Inputs and Outputs η Flow into the Policy π_i .

weight of the distance between policies. The origin of the set of policies, which is the vertex of the undirected graph, is $\pi_i \in \Pi, e \in E (e = \pi_i, \pi_j)$. The network structure of the graph is expressed using an adjacency matrix. The adjacency matrix is represented by a square matrix, and defines the elements in the matrix as follows: Fig. 2 shows an example of SAP-Net. A policy has able to multiple network path as output and input as shown in Fig. 3.

$$M_{ij} = \begin{cases} 0 & (e \notin E) \\ \omega_{ij} = 1 & (e \in E) \end{cases} \quad (3)$$

Shown in Fig. 2 by an adjacency matrix, the networks can be expressed as Eqn. (4). π_i shows the Qtable and an approximate model, which are the policies already acquired by the source-task

$$M = \begin{pmatrix} 0 & \omega_{12} & \omega_{13} & 0 \\ \omega_{12} & 0 & \omega_{23} & \omega_{24} \\ \omega_{13} & \omega_{23} & 0 & 0 \\ 0 & \omega_{24} & 0 & 0 \end{pmatrix} \quad (4)$$

When constructing SAP-Net, weights of each element are adjusted by influencing the weight of inter-policy distance in the prototype matrix generated from the classified categories. The prototype matrix is represented by an adjacency matrix like SAP-Net, and weight ω_c is generated from the connection between the policy distances in a given category, as shown in Eqn. (5).

$$M = M + \delta(\mathbb{P}_1 + \dots + \mathbb{P}_n) \quad (5)$$

n is the number of categories, and δ is a coefficient for adjusting the weight of the prototype matrix.

To calculate similarities and relevance of the policies, it is necessary to evaluate them from various viewpoints. In general, when making a judgment that A and B are related or similar, we count the number of related parts and consider that they are more similar as the number of related parts increases. In this research, we apply this idea and calculate relevance by adding up the prototype matrices generated with various prototypes.

E. Selection of Policy

To select the policy, the selection probability $P(\pi_i)$ of each policy is calculated based on the constructed SAP-Net and the activation value \mathbb{A}_i given to each policy. The calculation formula is shown in the following Eqn. (6). i, j are the numbers used to identify policies.

$$P(\pi_i) = \frac{\exp(\mathbb{A}_i)}{\sum_{j=0}^{n-1} \exp(\mathbb{A}_j)} \quad (6)$$

F. Activation Process

In this study, based on the loss $\mathbb{E}_t^{\pi_i}$ calculated when the agent acts by selecting a certain policy π , judgment of positive and negative transitions is made. The judgment formula is shown in Eqn. (7).

$$T_a = \mathbb{E}_{t+1}^{\pi_i} - \mathbb{E}_t^{\pi_i} \quad (7)$$

Here, T_a is threshold value for judgment. Based on the calculated judgment result, increase processing is applied to the activation value of the policy. The process applied to the activation value is shown in Eqn. (10). \mathcal{A} is a activated coefficient that adjusts the activation value increment. This process is called activation, and the activation value must be adjusted.

$$\mathbb{A}_i = \begin{cases} \mathbb{A}_i + \mathcal{A} & (T_a \leq 0) \\ \mathbb{A}_i & (T_a > 0) \end{cases} \quad (8)$$

G. Spreading Process

The spreading process of the activation value is applied to a policy with connection relation via policy distance extending from the activated policy. When defining the spreading value, the state of activation value η diffuses to the policy π stored on SAP-Net, as shown in Fig. 4. The activation value output is recorded when a certain policy is activated or diffused. The activation value output η_i from a certain policy π_i is obtained by the following expression using the number of policy distances extending from the policy π .

$$\eta_i \leftarrow \frac{\mathcal{A}}{k} \quad (9)$$

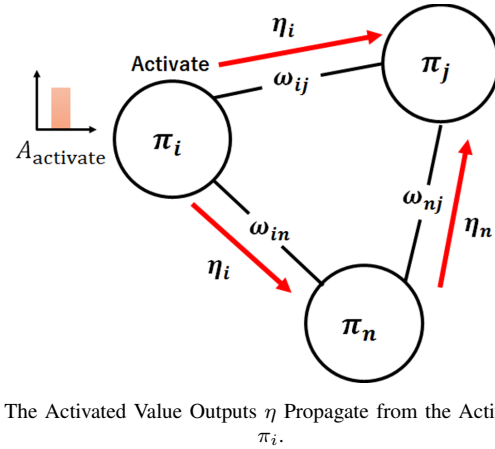


Fig. 4. The Activated Value Outputs η Propagate from the Activated Policy π_i .

Due to diffusion, the activation value output is updated every time via a policy. This is done based on Eqn. (9). Considering the diffusion of the activation value for the two sources of diffusion π_i and the diffusion destination π_j adjacent to the diffusion source, the active value change amount $\Delta\mathbb{A}_j$ takes into account the activation value output η_i from π_i and the diffusion through multiple policies. The sum $\sum \omega$ of the weights of the inter-policy distances by which it passed through, and h of inter-policy distances by which the intermediary passed through are obtained as follows.

$$\Delta\mathbb{A}_j = \begin{cases} 0 & (\sum \omega \geq T_\omega) \\ \frac{\eta_i}{h \sum \omega} & (\sum \omega < T_\omega) \end{cases} \quad (10)$$

An image of activation diffusion is shown in Fig. 4. Activation spreading diffuses permanently as long as it does not specify the range of spreading, because it is calculated recursively through each policy. Taking this into consideration, the threshold value T_ω is determined by the sum of the weights of the inter-policy distances by which the value passes through. Ultimately, each policy calculates the total sum of activation change amounts, as shown in Eqn. (11), by the number of active value outputs received through the network and adds it to the residual activation value. In the proposed method, we adjust the activation value of each policy through these processes, and re-learn while choosing a policy for each behavior. After the learning, the policy with the highest activation value is determined as a policy that can be learned effectively. The two algorithms describing the activation process and the spreading process of the proposed method are shown in Algorithm 1, 2.

$$\mathbb{A}_j = \begin{cases} \mathbb{A}_j + \sum \Delta\mathbb{A}_j & (T_a \leq 0) \\ \mathbb{A}_j & (T_a > 0) \end{cases} \quad (11)$$

Each policies are labelled environmental information of source task respectively. Therefore input of proposed system is environmental information of target task, and final output of proposed system is selected policy or policies.

Algorithm 1 Activation process

```
1: function ACTIVATION( $i, \mathbb{A}, M, \mathcal{A}$ )
2:   Initialize:  $h = k = j = \Omega = 0, N$  is set as number of policies collected via source-task
3:      $\triangleright h$  is inter-policy distances by witch the intermediary passed through            $\triangleright k$  is number of path that  $\pi_i$  has
4:      $\triangleright j$  is number of policy connected to  $\pi_i$  via  $d_{ij}$                                 $\triangleright \Omega$  is corresponding to  $\sum \omega$ 
5:    $\mathbb{A}_i \leftarrow \mathbb{A}_i + \mathcal{A}$ 
6:   for  $j = 0$  to  $N - 1$  do
7:     if  $M_{ij} \neq 0$  then
8:        $k \leftarrow k + 1$ 
9:     end if
10:  end for
11:   $j = 0$ 
12:  while  $j < N$  do
13:    if  $M_{ij} \neq 0$  then
14:       $\eta_i = \frac{\mathcal{A}}{k}$ 
15:       $\Omega \leftarrow \Omega + M_{ij}$ 
16:      SPREADING( $j, h, \Omega, \eta_i$ )
17:       $\Omega \leftarrow \Omega - M_{ij}$ 
18:    end if
19:     $j \leftarrow j + 1$ 
20:  end while
21: end function
```

Algorithm 2 Spreading process

```
1: function ACTIVATION( $j, h, \Omega, \eta_i$ )
2:   Initialize:  $k = 0$ 
3:    $\eta = \eta_i$ 
4:    $i = j$ 
5:    $h \leftarrow h + 1$ 
6:   if  $\Omega < T_\omega$  then
7:      $\mathbb{A}_i \leftarrow \mathbb{A}_i + \frac{\eta}{h\Omega}$ 
8:     for  $i = 0$  to  $N - 1$  do
9:       if  $M_{ij} \neq 0$  then
10:         $k \leftarrow k + 1$ 
11:       end if
12:     end for
13:      $j = 0$ 
14:     while  $j < N$  do
15:       if  $M_{ij} \neq 0$  then
16:         $\eta \leftarrow \frac{\eta}{k}$ 
17:         $\Omega \leftarrow \Omega + M_{ij}$ 
18:        SPREADING( $j, h, \sum \omega, \eta$ )
19:         $\Omega \leftarrow \Omega - M_{ij}$ 
20:       end if
21:     end while
22:   end if
23: end function
```

III. EVALUATION

A. Experimental Setup

This study conducted transition-based learning using the proposed method by establishing multiple approximate models that randomly set and learned hidden layers to verify the usefulness of it. The learning efficiency of reinforcement learning using this new transfer learning and normal Deep Q-Network was compared. In this experiment, we focused on the structure of the NN and discussed the learning effect when performing a selecting operation during re-learning. In

addition, we did not fix the weights of the policies reused at the time of transfer learning, and observed the learning effect when re-learning while selecting. In the case of not fixing the weights during transfer learning, convergence tends to be difficult. Therefore, whether it is necessary to fix weights when learning while selecting multiple policies was verified. In addition, the necessary conditions for adaptation to heterogeneous tasks and heterogeneous agents will be included in future prospects based on the learning results.

The source-task and the target-task were learned by the same task. In this experiment, the total reward obtained from the learning curve of each Episode versus the total reward, as well as the test results after learning, were evaluated. By this evaluation, we confirmed two points of effective learning and accurate learning.

B. Experimental Condition

For learning of the source-task and re-learning of the proposed method, a Deep Q-Network (DQN) built with library ChainerRL [19] was used. Experimental environment is build using Python, and for the learning task, CartPole and MountainCar, provided by OpenAI Gym [20], were adopted. Approximate models of the multiple methods used in the proposed method were designed randomly, assuming that a person cannot manually design an optimal hidden layer. The number of policies used in the proposed method was set to 10. In this experiment, the prototype used as a categorization perspective was set as the number of layers in the hidden layer of the policy.

For the DQN to be compared, the hidden layer was set to a 1 to 3 range, and the number of units per layer was set to 100. The number of units was approximately the same as the number of units in the hidden layer of the policy used for the proposed method. Statistic losses calculated for each step, which can be referenced by ChainerRL, were used for positive or negative judgment of policy transition. The learning

Algorithm 3 Transfer reinforcement learning with proposed method

Require: \mathbb{A} : Array that manages the activation value, M : SAP-Net, Π : Array that manages policies

- 1: N = is set as number of policies collected via source-task
- 2: $Episode = 1$
- 3: **for** $Episode = 1$ to Max Episode +1 **do**
- 4: $reward = 0$
- 5: $done = False$ ▷ $done$ Fragment of granting rewards (True or False)
- 6: $R = 0$
- 7: $step = 1$
- 8: $step_{store} = 1$ ▷ Fragment of store data for Experience Replay
- 9: **while** $done \neq True$ and $step < Max\ step$ **do**
- 10: **if** $step_{store} < N * \text{Replay start size}$ **then**
- 11: Act a and store data(s, a, s', r) to be used for Experience Replay when selecting each policy
- 12: **else**
- 13: Select policy π_i based on probability $P(\pi_i) = \frac{\exp(\mathbb{A}_i)}{\sum_{b=0}^{N-1} \exp(\mathbb{A}_b)}$
- 14: Act a and training with π_i
- 15: Calculates statistic loss $\mathbb{E}_t^{\pi_i}$
- 16: $T_a = \mathbb{E}_{t+1}^{\pi_i} - \mathbb{E}_t^{\pi_i}$ ▷ Judgment positive or negative of policy transition
- 17: **if** $T_a > 0$ **then** SPREAING($i, \mathbb{A}, M, 0$) ▷ Negative transfer
- 18: **else** SPREAING($i, \mathbb{A}, M, \mathcal{A}$) ▷ Positive transfer
- 19: **end if**
- 20: $R \leftarrow R + r$
- 21: $step \leftarrow step + 1$
- 22: $step_{store} \leftarrow step_{store} + 1$
- 23: **end if**
- 24: **end while**
- 25: $step = 1$
- 26: $step_{store} = 1$
- 27: **end for**

TABLE I. TASK AND PARAMETER SETTING

Learning task	CartPole	MountainCar
Max episode	5000	10000
Number of test	5	5
Activate function	ReLU	ReLU
Discount rate	0.99	0.99
Explorer	ϵ -greedy ($\epsilon = 0.03$)	ϵ -greedy ($\epsilon = 0.03$)
Optimizer	Adam ($\epsilon = 1e - 3$)	Adam ($\epsilon = 1e - 3$)
Replay buffer size	$1e + 6$	$1e + 6$
Replay start size	50	100

TABLE II. PARAMETER SETTING FOR PROPOSED METHOD

Parameter	Symbol	Value
Adjustment weight of \mathbb{P}	δ	-0.5
Initial weight of \mathbb{P}	ω_c	0.1
Activation function parameter	\mathcal{A}	1.0
Threshold	T_w	1.0

conditions of DQN are shown in Table I. For parameters other than Table I, ChainerRL's initial setting was used. Episode per total rewards was averaged 5 times, test results were summarized 5 times and expressed as a learning curve and bar graph. The learning curves are indicated by the moving average value every 10 episodes in order to easily observe the increase or decrease of the total reward for each episode. Table II shows the parameters of the proposed method. The transfer learning algorithm of reinforcement learning used in the proposed method is shown in Algorithm 3.

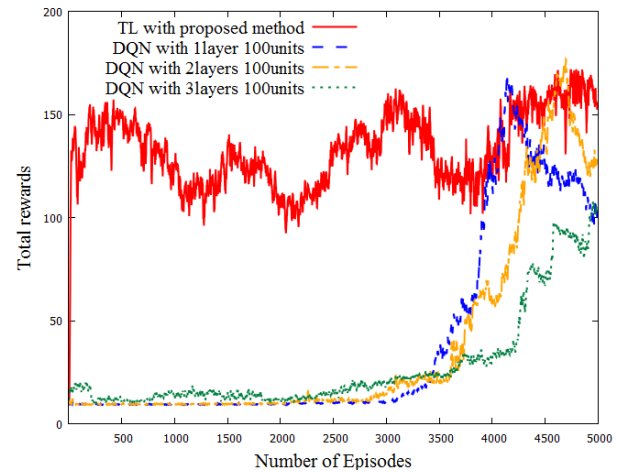


Fig. 5. Comparison with Learning Curve (CartPole).

C. Experimental Results

The learning curve of the CartPole task is shown in Fig. 5, and the test result after learning is shown in Fig. 6. The learning curve of the MountainCar task is shown in Fig. 7, and the test result after learning is shown in Fig. 8. In Fig. 5 and Fig. 7, the red learning curve shows transfer learning (TL: Transfer Learning) using the proposed method. The blue learning curve represents the result of reinforcement learning (DQN: Deep Q-Network) of 100 units in one layer, the orange

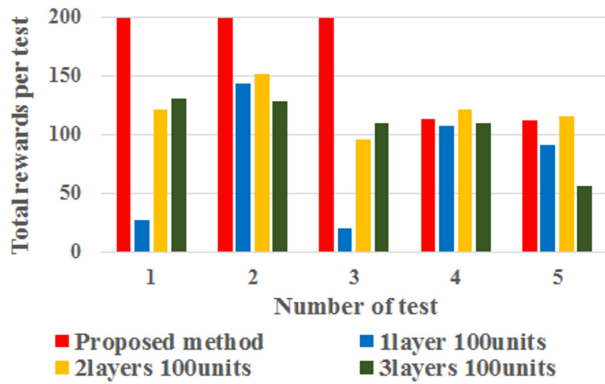


Fig. 6. Test Task After Learning (CartPole).

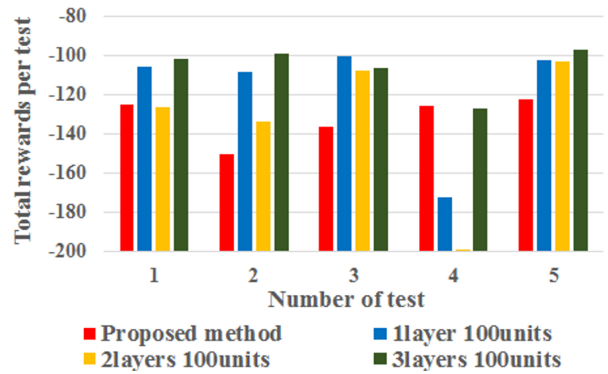


Fig. 8. Test Task After Learning (MountainCar).

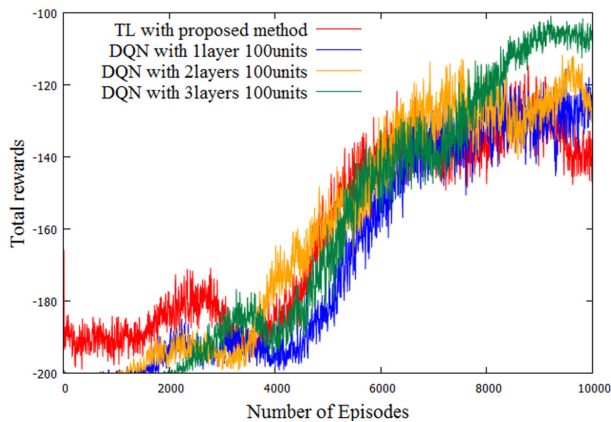


Fig. 7. Comparison with Learning Curve (MountainCar).

learning curve of 100 units in 2 layers, and the green learning curve of 100 units in three layers. The vertical axis represents the total rewards, and the horizontal axis represents the task execution number (Episode).

In Fig. 6 and Fig.8, the vertical axis represents the total rewards, and the horizontal axis represents the number of tests. The color of the bar graph shown in the test result corresponds to the color of the learning curve.

From the results shown in Fig. 5, it can be confirmed that high rewards can be earned from an Episode at the initial learning stage. From this result, it is understood that the learning is promoted by the proposed method. From the learning results of the three types of DQN to be compared, a phenomenon in which the reward acquisition amount declined was seen. This may be caused by the fact that the structure of the learning model is not optimized for the task. When learning from the beginning, in order to perform optimum learning, it is necessary to adjust the NN structure and its parameters to a specific task. From the test results of CartPole in Fig. 6, the proposed method confirmed that it is possible to acquire high rewards at high frequency. This is presumed to be due to re-learning of the policy with the highest priority being possible while selecting the policy. We confirmed that the reward decreased in the 4th and 5th tests. It is presumed that this was caused by the fluctuation of the teacher data value during the learning process. In this regard, it may be possible

to cope with by fixing the weight of the approximate model for each policy. From the above results, it was possible to confirm the usefulness of the proposed method in CartPole.

From the results shown in Fig. 7, it can be confirmed that, although the amount of reward earned at the early stage of learning was large, the proposed method decreased the amount of compensation earned in the Episode near the end of learning, compared to the DQN. This result seems to be caused by being unable to perform the MountainCar task. In this task, judging positive or negative transition is carried out at the end of the task, while in the proposed method, the transfer judgment of the policy is carried out for each action. In order to increase the reward acquisition amount against this result, this may be solved by examining whether to perform the policy transition judgment for each Episode or to perform it every certain step number. In addition, by considering indices other than the loss for activation judgment, some conditions may be judged more effectively.

Considering the usable range of the proposed method, we think that it can be applied to tasks where timing of reward assignment is likely to be associated with task achievement condition. In future prospects, if SAP-Net includes policies for heterogeneous tasks and agents, we estimate that a new method is needed that takes into account the above applicable range. For that method, it is desirable to calculate selection candidates of policies by using categories, or to exclude policies presumed to be unnecessary for learning of the target-task. In the case of policies in heterogeneous agents, processing the data to enable reusability by the target-task through Inter task mapping is necessary.

IV. CONCLUSION

In this paper, we proposed a method to re-learn while choosing a policy as transfer learning. The method was implemented by using a spreading activation model and was verified by a computer experiment. From the results of the verification, usefulness was suggested in the transfer learning of the same task without manual evaluation by NN model design. In addition, as a countermeasure to the problem, we will examine judgment of appropriate positive and negative transition, and consider selection candidate calculation of policies using categories in transfer learning of heterogeneous tasks and agents. We also think that it is possible to develop

a method of applying Inter task mapping for related tasks by applying categories to this problem.

ACKNOWLEDGMENT

Part of this research was undertaken with the aid of JSPS Grant-in-Aid for Scientific Research (JP16K12493 and JP19K12173). We express our gratitude here.

REFERENCES

- [1] T. Tongloy, S. Chuwongin, K. Jaksukam, C. Chousangsunton, and S. Boonsang, "Asynchronous deep reinforcement learning for the mobile robot navigation with supervised auxiliary tasks," in *2017 2nd International Conference on Robotics and Automation Engineering (ICRAE)*, 2017, pp. 68–72.
- [2] T. Shimizu, R. Saegusa, S. Ikemoto, H. Ishiguro, and G. Metta, "Robust sensorimotor representation to physical interaction changes in humanoid motion learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1035–1047, 2015.
- [3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, 1998.
- [4] M. E. Taylor, *Transfer in Reinforcement Learning Domains*, ser. Studies in Computational Intelligence. Springer, 2009, vol. 216.
- [5] F. Fernández and M. Veloso, "Learning domain structure through probabilistic policy reuse in reinforcement learning," *Progress in Artificial Intelligence*, vol. 2, no. 1, pp. 13–27, 2013.
- [6] T. Takano, H. Takase, H. Kawanaka, and S. Tsuruoka, "Transfer method for reinforcement learning in same transition model – quick approach and preferential exploration," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 1, 2011, pp. 466–469.
- [7] H. Kono, R. Katayama, Y. Takakuwa, W. Wen, and T. Suzuki, "Activation and spreading sequence for spreading activation policy selection method in transfer reinforcement learning," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 7–16, 2019.
- [8] A. M. Collins, F. Elizabeth, and Loftus, "A spreading-activation theory of semantic processing," *Psychological Review*, vol. 82, no. 6, pp. 407–428, 1975.
- [9] M. Han and B. Zhang, "Control of robotic manipulators using a cmac-based reinforcement learning system," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, vol. 3, 1994, pp. 2117–2122.
- [10] Y.-P. Hsu, W.-C. Jiang, and H.-Y. Lin, "A cmac-q-learning based dyna agent," in *2008 SICE Annual Conference*, 2008, pp. 2946–2950.
- [11] F. Li, F. Luo, Y. Gao, D. Qi, and J. Hu, "Research on fuzzy reinforcement learning algorithm for agents in grids," in *2009 Third International Symposium on Intelligent Information Technology Application Workshops*, 2009, pp. 336–339.
- [12] X.-N. Wang, X. Xu, and H.-G. He, "Policy gradient fuzzy reinforcement learning," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, vol. 2, 2004, pp. 992–995.
- [13] H. S. Cho, "A study on the control of nonlinear system using growing rbf and reinforcement learning," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 5, 2007, pp. 521–525.
- [14] S. Li, L. Ding, H. Gao, Y. Liu, N. Li, and Z. Deng, "Reinforcement learning neural network-based adaptive control for state and input time-delayed wheeled mobile robots," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4171–4182, 2018.
- [15] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 81–85.
- [16] 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, Feb. 2015.
- [17] T. Okuyama, T. Gonsalves, and J. Upadhyay, "Autonomous driving system based on deep q learnig," in *2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2018, pp. 201–205.
- [18] H. Sasaki, T. Horiuchi, and S. Kato, "A study on vision-based mobile robot learning by deep q-network," in *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 2017, pp. 799–804.
- [19] Chainerrl. [Online]. Available: <https://github.com/chainerrl/chainerrl>
- [20] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016, cite arxiv:1606.01540. [Online]. Available: <http://arxiv.org/abs/1606.01540>

The Adoption of Mobile Health Applications by Patients in Developing Countries: A Systematic Review

Nasser Aljohani¹, Daniel Chandran²

Faculty of Engineering and Information Technology
University of Technology Sydney Australia, Sydney, Australia

Abstract—Mobile health (m-health) apps adoption in developing countries is a new research area in the healthcare industry. M-health is comparatively recent in information systems, with little attention being paid to it developing countries in the previous years. Applications of the m-health strategies in developing nations are considered one of the best platforms for guaranteeing the citizenry's safety and healthcare security. A systematic review was conducted of m-health apps adoption by patients in developing countries to evaluate the current results. It reviews 22 papers that were published on the topic of m-health adoption in developing countries in academic journals and conferences over the last decade. It identifies the research in terms of research methodologies, theories and models adopted, significant factors identified, limitations and recommendations. Findings show there is a limited contribution to m-health apps adoption in developing countries. Most studies employed TAM and focused on the technological and individual levels; very low intention has been made to health-related factors, levels, and theories. The review presents a broad overview of previous academic studies with a view to future research.

Keywords—M-health; mobile health; apps; adoption; review; developing countries

I. INTRODUCTION

Mobile technology has grown in use in the healthcare delivery and health results in developed countries in the last few years. Electronic health (e-health) refers to computer-based services, while mobile health (m-health) applies to mobile systems with specialized features to improve health care delivery [5]. The concept "m-health" was first used by Prof. Isteparian, who refers to mobile devices and networking systems used for healthcare delivery [28]. Fig. 1 presents the role of both e-health and m-health.

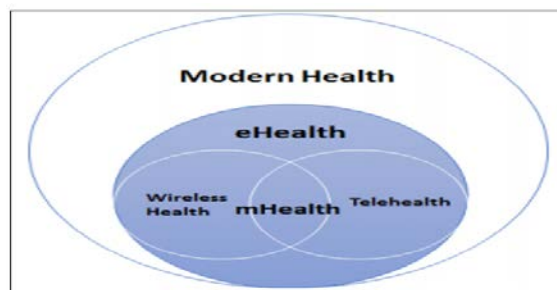


Fig. 1. Role of E-health and M-health (Adopted from [26]).

In Fig. 2, Dehzad et al. classify m-health into three classes of technologies: devices, sensors, and applications [10]. Besides, they categorized the target group of m-health into healthy people, hospital patients, and chronically ill individuals. M-health solutions can carry four areas: wellness and prevention, diagnosis, treatment and [10]. There are currently more than 165,000 health apps available on smartphone online stores [9]. According to Larson [29], m-health apps categorized by searchers into four different types:

- Information app: Provide general health information to the public.
- Diagnostic app: To enter patients' information and provide a diagnosis to physicians.
- Control app: Assist medical devices with remote monitoring and control.
- Adapter app: Transform smartphones to become mobile medical devices.

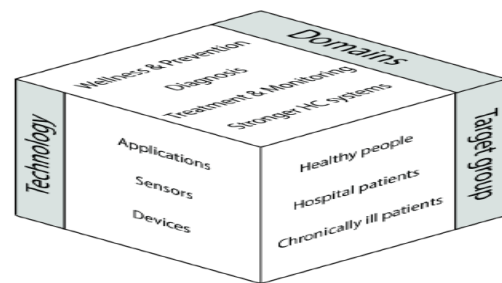


Fig. 2. Classification Model for M-health Solutions (Adopted from [10]).

Having m-health technology will promote healthcare awareness and access to knowledge that will enhance the lives of the citizenry and provide the developed countries with the opportunity to create a sustainable workforce with economic resilience [16]. The majority of mobile phone users, especially smartphones and computing technology in developing nations, have adopted mobile devices to access information, particularly about the right healthcare facilities that offer quality services to guarantee security and healthcare safety. The initiative of downloading mobile apps for understanding the proper treatment procedures and the most appropriate types of medication has enabled the majority of the citizens in the developing nations to access Medicare conveniently and reduce cost [1]. The adoption of m-health gives the governments of

the developing nations the advantage to guarantee patients safety and health information records by considering data protection and visualization strategies. Some developing countries, such as Malaysia, Thailand, China, and India, have been introducing technological advances in healthcare systems to improve the treatment process [24].

Despite the prospective benefits of m-health apps, adopting and accepting such a technology is not as widespread as expected in developing countries [30]. Factors found to be influencing m-health have been explored in several studies. However, an insufficient contribution has been paid to the factors affecting the adoption of m-health by patients; besides, the factors influencing e-health adoption rates in developed countries have been comprehensively reviewed against m-health. In the sense of developing countries, we conducted a systematic literature review to properly comprehend and verify the adoption of m-health apps. This is an especially significant discovery in the advancement of a modern research area. It provides a possibility to step back and review several samples, methods, and theories collected from different studies in m-health. Thus, this research seeks to play a key role in enhancing the research in this rapidly growing field of m-health. Beyond that to examine the existing state of m-health among people of devolving countries.

II. LITERATURE SEARCH APPROACH

Several keyword sets were evaluated to have a sufficiently reliable and authentic secondary source. For instance, the sets of keywords considered for the study include "adoption of m-health", "adoption of mobile health", and "adoption of Google mhealth" and this focused on the use and application of Google Scholar as one of the most reliable search engines for academic journals with both comprehensive and conclusive information about the initiative of developing the m-health in the developing nations. The technique that was idealized to establish the search setting considered extensive and varied studies published between 2010 and 2020. The core underlying principle for choosing the wide range of timeline is to present information with a broad historical background about the nature of the developing countries' approach to adopting m-health.

The major reasoning for choosing the wide range of timelines is dependent on two fundamental reasons. First, due to the nature and the quality of healthcare outcomes before and after the developing nations begun the initiative of adopting the m-health. It is imperative to note that restricting the healthcare systems in developing countries by adopting the m-health strategy has improved and stands a better chance to improve the citizens' overall quality of lifestyle in the developing nations in the foreseeable future [25]. Second, by 2013, the use and application of mobile technology to access and receive healthcare services had become most popular since both the government and the citizenry in the developing nations considered the strategy one of the best platforms of accessing quality healthcare conveniently and at a reduced cost [14]. The use of mobile apps can be a fast and reliable way to provide healthcare to large numbers of people who cannot make a physical visit to a healthcare facility.

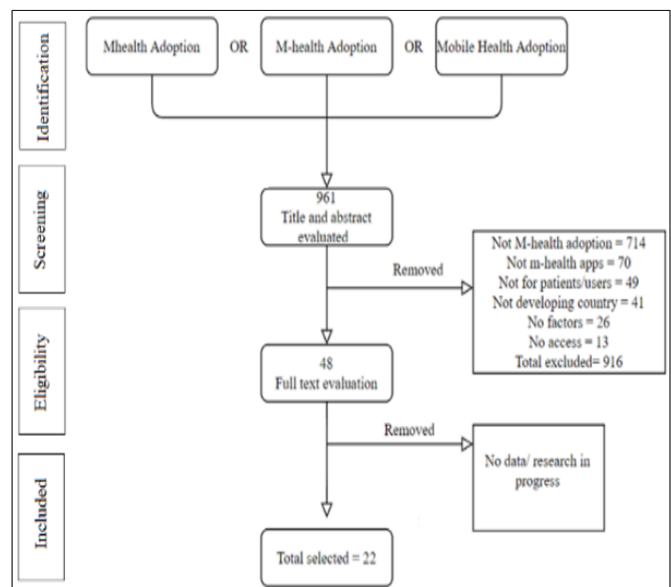


Fig. 3. Flowchart of the Search Strategy.

A total of 22 secondary sources were selected for the study, particularly academic journals, to investigate the approach taken by the developing nations for reasons of adopting the m-health strategies. These studies proved relevant and authentic about affirming the perspective that the developing countries should adopt the m-health approach to improve the overall quality of lifestyle and health condition of their citizens. Using the PRISMA flow diagram [32], Fig. 3 shows the research selection strategy.

III. LITERATURE ANALYSIS

This section explains several of the existing primary literature concerning m-health from a scholarly perspective. The analysis included the date of publication, theories and methods used the identified significant factors, and the limitations and findings of the m-health applications research.

A. General Analysis

1) *Publication date analysis*: Publication dates were analyzed to determine the latest most significant developments in research publications. As shown in Fig. 4, the number of articles increased from 2012 until 2015. Then, the number increased to reach 3 to 4 papers during the years 2016 to 2018. No complete studies have been conducted in developing countries during the last two years. This could be because m-health concept is new or has not been applied thoroughly in developing countries.

2) *Countries and number of participants analysis*: The analysis of studies conducted in developing countries showed that ten studies were conducted in China, five studies in Bangladesh, two in Jordan, one in Arab countries, one in Taiwan, one in Malaysia, and one in United Arab Countries. Table I and Fig. 5 show broad information about each study.

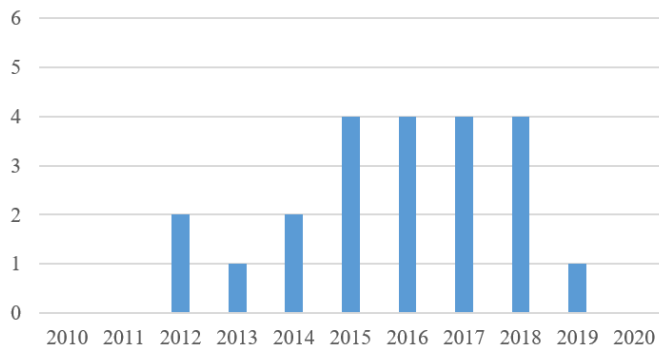


Fig. 4. Number of Studies Per Years (2010-2020).

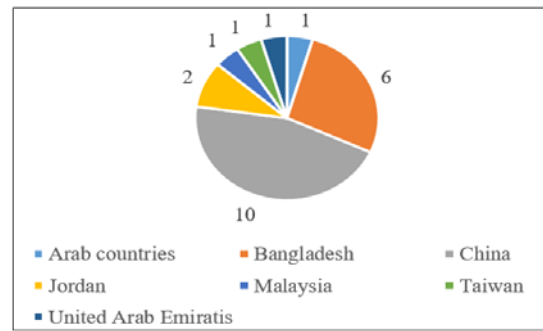


Fig. 5. Number of Studies Per Countries.

TABLE I. DETAILS OF THE STUDIES

Country	Targeted group	Participants	Reference
Arab countries	Elderly Arab m-health users	134	[7]
Bangladesh	Young citizen of public and private universities	144	[20]
Bangladesh	Elderly users	375	[14]
Bangladesh	All	227	[21]
Bangladesh	Patients	37	[27]
Bangladesh	Elderly users above 60	274	[22]
Bangladesh	All	296	[3]
China	All	429	[17]
China	Elderly users	212	[33]
China	M-health users Over 40	424	[11]
China	All	481	[35]
China	M-health users	428	[18]
China	All	650	[17]
China	Hypertensive patients	157	[13]
China	M-health service users	650	[36]
China	All	388	[12]
China	Elderly users above 60	395	[31]
Jordan	All	366	[15]
Jordan	All	365	[2]
Malaysia	All	480	[24]
Taiwan	Young users	170	[23]
United Arab Emirates	M-health users	144	[6]

B. Theories and Models Used

Different theories and model have been used in the 22 studies. These include the Technology and Acceptance Model, Theory of Reasoned Action (TRA), Unified Theory of Acceptance and Use (UTAUT), Protection Motivation Theory (PMT), Theory of Planned Behavior (TPB). It has been noted that the TAM model is the most applied model among the 22 selected papers. Some researchers used the same model, extended, or even combined it with other models. Two studies proposed hypotheses without specifying any theory or model. In addition, two studies have suggested new models. However, one study, which applied quantitative methodology, did not involve any theory. Table II shows the theories used in each study and their references.

C. Research Methodologies

The most popular research approach used in the m-health area is the quantitative research technique, while qualitative research is used only by Khatun et al. [27] out of the 22 studies. This could be due to the impossibility to interview patients directly. Health conditions of patients could be the possible reason.

D. Significant Factors

This section identifies only factors that have been proven to be affecting the behavioral intention in different countries. It has been noted the most significant factor identified in most studies is the preserved ease of use. The following Table III shows the significant factors, the number of studies, and their references.

TABLE II. THEORIES AND MODELS USED

Theory/ Model	Number of studies	Reference
TAM	10	[6] [7] [13] [17] [20] [21] [33] [15] [23] [36]
UTAUT	5	[2] [3] [14] [22] [33]
TRA	2	[33] [35]
PMT	2	[18] [33]
TPB	2	[33] [11]
New model	2	[17] [31]
No Model	2	[24] [27]
VAB	1	[11]

TABLE III. IDENTIFIED SIGNIFICANT FACTORS

Factor	Reference	N	Factor	Reference	N
Technical					
Perceived ease of use	[2] [6] [7] [12] [13] [15] [20] [21] [23] [33]	10	Perceived usefulness	[2] [6] [12] [15] [20] [21] [23] [33]	8
Performance expectancy	[3] [14] [22] [33]	4	Facilitating conditions	[3] [14] [35]	3
Effort expectancy	[3] [14] [22]	3	Resistance to change	[13] [22]	2
Technology anxiety	[22]	1	Technological incapability	[27]	1
Lack of access	[27]	1	Technological incapability	[27]	1
Individual					
Attitude	[7] [23] [24] [33] [35]	6	Trust	[6] [12] [15] [19] [27] [31]	6
Age	[18] [19] [23]	3	Self-Efficacy	[13] [18] [33]	3
Response Efficacy	[18] [33]	2	Awareness	[2] [27]	2
Gender differences	[27] [35]	2	Innovativeness	[2]	1
Smartphone technology usage experience	[13]	1	Illiteracy	[27]	1
perceived reliability	[3]	1	Language	[27]	1
Perceived Behavioral Control	[33]	1	Poverty	[27]	1
Waiting time	[14]	1	Perceived personalization	[19]	1
Social/Cultural/Environmental					
Social influence	[2] [3] [13] [14] [15] [22]	6	Subjective norm	[21] [33] [35]	3
Culture	[7]	1			
Security/Privacy					
Privacy	[12] [15] [19]	3	Security	[6] [15]	2
Performance risks	[12]	1			
Health					
Perceived Severity	[18] [33]	2	Perceived Vulnerability	[18] [33]	2
Relationship with the doctor	[13]	1	Support from hospital	[31]	1
Chronic disease	[12]	1	Declining physiological conditions	[31]	1

IV. DISCUSSION OF RESULTS AND LIMITATIONS

The number of studies in m-health adoption in developing countries is still low compared to the total number of developing nations. There is an increased interest and attention by researchers in China, 10 studies, and Bangladesh, 6 studies, about the adoption of m-health apps by patients in the last ten years. However, the number of studies comparing to the number of developing countries is still low. Previous literature analysis revealed that little to no attempt to exploit qualitative or mixed methods had been made by present researchers. Only one research out of the 22 papers has used the qualitative approach. The qualitative method in healthcare can answer difficult questions or questions that may not be answered by quantitative research considering the context in which it has been examined [4]. Since mixed study methodology affords a clearer perspective and further interpretations of the conclusions and can provide richer account of healthcare than any approach can provide alone [34], both qualitative and mixed methodologies are crucially needed in m-health studies.

Regarding theories and models applied to m-health adoption studies, the TAM model has been mostly used, followed by UTAUT over other theories. The selected papers mostly focused on the technological and individual levels with the minimal intention to health-related factors. M-health is a combination of technology and health areas. Only three studies have mentioned some health-related factors in their studies, the role of chronic disease [12], declining physiological conditions [31], and relationship with doctors [13]. Some studies have used the theories with no consideration of the context that had been examined. Meng et al. [31] state a limitation of their study that it may not be applicable to other countries due to cultural differences. Chandran and Aljohani state that Saudi Arabian culture, for example, is a mixture of both traditions and Islamic beliefs and call for more consideration [8]. This is also applicable to other Arab and Islamic nations. From the analysis of the selected 22 studies, only one study considered cultural affect. Hence, there is a need to put more efforts to examine health-related factors by considering cultural aspects.

Age has been considered as a targeted group in some studies. For example, elderly m-health users have been included as the main targeted group in five studies [7-14-22-33-31] and young m-health users in two studies [20-23]. It would be more useful for future contributions to consider age to be a moderator to target more participants instead of specifying the sample size. Only two studies have examined the role of gender as a significant factor. Hence, there is a need to put some efforts into age and gender as moderators of factors affecting the adoption of m-health apps in developing countries.

In summary, there is a call for more studies about m-health adoption in developing countries. Exploring the adoption of m-health by applying qualitative or mixed methods will yield more excellent perspectives and offer more reliable evidence to m-health apps studies. There is more space for future research to analyze the impact of m-health on health and cultural factors.

V. CONCLUSION

A systematic review of literature to evaluate at m-health app adoption in developing countries between 2010 and 2020 was conducted for this study. Among 48 studies, 22 studies were included in the review, thereby proving to be suitable. Most studies had used quantitative methodology, but this particular one chose the qualitative method, and no one attempted to employ a mixed method. As most studies used the TAM model and focused on the technological and individual levels, the very low intention has been made to other health-related factors, levels, and theories. Moreover, there is a lack to consider the culture being examined. The review presents a broad overview of previous academic studies with a view to future research. This study would be useful as guide to other researchers in the future. Lastly, another research field for e-health is to look into whether e-health research can also be extended to m-health.

REFERENCES

- [1] A. Alaiad, M. Alsharo, and Y. Alnsour, "The Determinants of M-Health Adoption in Developing Countries: An Empirical Investigation," *Applied clinical informatics*, vol. 10, no. 5, pp. 820-840, 2019.
- [2] A. Alalwan, A. M. Baabdullah, N. P. Rana, Y. K. Dwivedi, F. Hudaib, and A. Shammout, "Examining the Factors Affecting Behavioural Intention to Adopt Mobile Health in Jordan," in *Conference on e-Business, e-Services and e-Society*, Springer, pp. 459-467, 2018.
- [3] M. Z. Alam, W. Hu, and Z. Barua, "Using the UTAUT model to determine factors affecting acceptance and use of mobile health (mHealth) services in Bangladesh," *Journal of Studies in Social Sciences*, vol. 17, no. 2, 2018.
- [4] Z. Q. Al-Busaidi, "Qualitative research and its uses in health care," *Sultan Qaboos University Medical Journal*, vol. 8, no. 1, p. 11, 2008.
- [5] N. Aljohani and D. Chandran, "Adoption of M-Health Applications: The Saudi Arabian Healthcare Perspectives," in *Australasian Conference on Information Systems*, Perth Western Australia, AIS eLibrary, pp. 180-186, 2019.
- [6] M. Alloghani, A. Hussain, D. Al-Jumeily, and O. Abuelma'atti, "Technology Acceptance Model for the Use of M-Health Services among health related users in UAE," in *2015 International Conference on Developments of E-Systems Engineering (DeSE)*, IEEE, pp. 213-217, 2015.
- [7] A. Alsswey, I. Naufal, and B. Bervell, "Investigating the Acceptance of Mobile Health Application User Interface Cultural-Based Design to Assist Arab Elderly Users," *International Journal of Advanced Computer Science and Applications*, vol. 9, doi: 10.14569/IJACSA.2018.090819, 2018.
- [8] D. Chandran and N. Aljohani, "The Role of Cultural Factors on Mobile Health Adoption: The Case of Saudi Arabia," *AMCIS 2020 Proceedings* 3, 2020.
- [9] C. Crico, C. Renzi, N. Graf, A. Buyx, H. Kondylakis, L. Koumakis, and G. Pravettoni, "mHealth and telemedicine apps: in search of a common regulation," *ecancermedalscience*, vol. 12, 2018.
- [10] F. Dehzad, C. Hilhorst, C. de Bie, and E. Claassen, "Adopting health apps, what's hindering doctors and patients?," *Health*, vol. 6, no. 16, p. 2204, 2014.
- [11] Z. Deng, X. Mo, and S. Liu, "Comparison of the middle-aged and older users' adoption of mobile health services in China," *International journal of medical informatics*, vol. 83, no. 3, pp. 210-224, 2014.
- [12] Z. Deng, Z. Hong, C. Ren, W. Zhang, and F. Xiang, "What predicts patients' adoption intention toward mHealth services in China: empirical study," *JMIR mHealth and uHealth*, vol. 6, no. 8, p. e172, 2018.
- [13] K. Dou, P. Yu, N. Deng, F. Liu, Y. Guan, Z. Li, Y. Ji, N. Du, X. Lu, and H. Duan, "Patients' Acceptance of Smartphone Health Technology for Chronic Disease Management: A Theoretical Model and Empirical Test," *JMIR mHealth and uHealth*, vol. 5, no. 12, pp. e177, 2017.
- [14] Y. K. Dwivedi, M. A. Shareef, A. C. Simintiras, B. Lal, and V. Weerakkody, "A generalised adoption model for services: A cross-country comparison of mobile health (m-health)," *Government Information Quarterly*, vol. 33, no. 1, pp. 174-187, 2016.
- [15] K. M. Faqih, and M.-I. R. M. Jaradat, "Mobile healthcare adoption among patients in a developing country environment: Exploring the influence of age and gender differences," *International Business Research*, vol. 8, no. 9, pp. 142, 2015.
- [16] M.-P. Gagnon, P. Ngangue, J. Payne-Gagnon, and M. Desmarts, "m-Health adoption by healthcare professionals: a systematic review," *Journal of the American Medical Informatics Association*, vol. 23, no. 1, pp. 212-220, 2016.
- [17] X. Guo, J. Yuan, X. Cao, and X. Chen, "Understanding the acceptance of mobile health services: A service participants analysis," in *2012 International Conference on Management Science & Engineering 19th Annual Conference Proceedings*, IEEE, pp. 1868-1873, 2012.
- [18] X. Guo, X. Han, X. Zhang, Y. Dang, and C. Chen, "Investigating m-health acceptance from a protection motivation theory perspective: gender and age differences," *Telemedicine and e-Health*, vol. 21, no. 8, pp. 661-669, 2015.
- [19] X. Guo, X. Zhang, and Y. Sun, "The privacy-personalization paradox in mHealth services acceptance of different age groups," *Electronic Commerce Research and Applications*, vol. 16, pp. 55-65, 2016.
- [20] M. R. Hoque, M. R. Karim, and M. B. Amin, "Factors affecting the adoption of mHealth services among young citizen: A Structural Equation Modeling (SEM) approach," *Asian Business Review*, vol. 5, no. 2, pp. 60-65, 2015.
- [21] M. R. Hoque, "An empirical study of mHealth adoption in a developing country: the moderating effect of gender concern," *BMC medical informatics and decision making*, vol. 16, no. 1, p. 51, 2016.
- [22] R. Hoque and G. Sorwar, "Understanding factors influencing the adoption of mHealth by the elderly: An extension of the UTAUT model," *International journal of medical informatics*, vol. 101, pp. 75-84, 2017.
- [23] M.-C. Hung and W.-Y. Jen, "The adoption of mobile health management services: an empirical study," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1381-1388, 2012.
- [24] Z. Hussein, W. Oon, and A. Fikry, "Consumer Attitude: Does It Influencing the Intention to Use mHealth?," *Procedia Computer Science*, vol. 105, pp. 340-344, doi: 10.1016/j.procs.2017.01.231, 2017.
- [25] B. Hwabamungu and Q. Williams, "m-Health adoption and sustainability prognosis from a care givers' and patients' perspective," in *Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, 2010, pp. 123-131.
- [26] International Pharmaceutical Federation (IPF), Use of mobile health tools in pharmacy practice, May 1, 2019. Accessed on: December 20, 2020. [Online]. Available: <https://www.fip.org/files/content/publications/2019/mHealth-Use-of-mobile-health-tools-in-pharmacy-practice.pdf>

- [27] F. Khatun, A. E. Heywood, P. K. Ray, A. Bhuiya, and S.-T. Liaw, "Community readiness for adopting mHealth in rural Bangladesh: a qualitative exploration," *International journal of medical informatics*, vol. 93, pp. 49-56, 2016.
- [28] E. G. Kariuki and P. Okanda, "Adoption of m-health and usability challenges in m-health applications in Kenya: Case of Uzazi Poa m-health prototype application," in *2017 IEEE AFRICON*, IEEE, pp. 530-535, 2017.
- [29] R. S. Larson, "A Path to Better-Quality mHealth Apps," *JMIR mHealth and uHealth*, vol. 6, no. 7, p. e10414, 2018.
- [30] L. Lee and A. Sheikh, "Understanding stakeholder interests and perspectives in evaluations of health IT," *Evidence-Based Health Informatics: Promoting Safety and Efficiency Through Scientific Methods and Ethical Policy*, vol. 222, p. 53, 2016.
- [31] F. Meng, X. Guo, Z. Peng, K.-H. Lai, and X. Zhao, "Investigating the Adoption of Mobile Health Services by Elderly Users: Trust Transfer Model and Survey Study," *JMIR mHealth and uHealth*, vol. 7, no. 1, p. e12269, 2019.
- [32] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement". *Open Med*, 3(3); 123-130, 2009.
- [33] Y. Sun, N. Wang, X. Guo, and Z. Peng, "Understanding the acceptance of mobile health services: a comparison and integration of alternative models," *Journal of Electronic Commerce Research*, vol. 14, no. 2, p. 183, 2013.
- [34] J. P. Wisdom, M. A. Cavaleri, A. J. Onwuegbuzie, and C. A. Green, "Methodological reporting in qualitative, quantitative, and mixed methods health services research articles," *Health services research*, vol. 47, no. 2, pp. 721-745, 2012.
- [35] X. Zhang, X. Guo, K.-h. Lai, F. Guo, and C. Li, "Understanding gender differences in m-health adoption: a modified theory of reasoned action model," *Telemedicine and e-Health*, vol. 20, no. 1, pp. 39-46, 2014.
- [36] X. Zhang, X. Han, Y. Dang, F. Meng, X. Guo, and J. Lin, "User acceptance of mobile health services from users' perspectives: The role of self-efficacy and response-efficacy in technology acceptance," *Informatics for Health and Social Care*, vol. 42, no. 2, pp. 194-206, 2017.

Method for Most Appropriate Plucking Date Determination based on the Elapsed Days after Sprouting with NIR Reflection from Sentinel-2 Data

Kohei Arai¹

Faculty of Science and Engineering
Saga University, Saga City
Japan

Yoshiko Hokazono²

Oita Prefectural Agriculture, Forestry and Fisheries
Research Center, Bungo-Ohno City
Oita, Japan

Abstract—Method for most appropriate plucking date determination based on the elapsed days after sprouting with Near Infrared: NIR reflection from Sentinel-2 data is proposed. Depending on the elapsed days after sprouting, tealeaf quality is decreasing. On the other hand, tealeaf yield is increasing with increasing of the days after sprouting. Therefore, there is most appropriate plucking date is very important. Usually, it is determined by the normalized Difference Vegetation Index: NDVI derived from handheld NDVI cameras, drone mounted NDVI cameras, and visible to NIR radiometer onboard satellites because NIR reflection and NDVI depend on tealeaf quality and yield. It, however, does not work well in terms of poor regression performance and species dependency. Moreover, it takes time consumable works for finding appropriate tealeaves from the acquired camera images. The proposed method uses only the days after sprouting. Next thing it has to do is to determination of sprouting date. In order to determine the date, optical sensor onboard Sentinel-2 data is used. Through experiment with the truth data taken at the intensive study area of the Oita Prefectural Agriculture, Forestry and Fisheries Research Guidance Center: OPAFFRGC, it is found that the proposed method is validated.

Keywords—Plucking date; elapsed days after sprouting; NIR reflection; sentinel-s; normalized difference vegetation index: NDVI

I. INTRODUCTION

Vegetation monitoring is attempted with red and photographic cameras [1]. Growth rate monitoring is also attempted with spectral observation [2].

Total nitrogen content corresponds to amid acid which is highly correlated to Theanine: 2-Amino-4-(ethyl carbamoyl) butyric acid for tealeaves so that total nitrogen is highly correlated to tea taste. Meanwhile fiber content in tealeaves has a negative correlation to tea taste. Near Infrared: NIR camera data shows a good correlation to total nitrogen and fiber contents in tealeaves so that tealeaves quality can be monitored with network NIR cameras.

It is also possible to estimate total nitrogen and fiber contents in leaves with remote sensing satellite data, in particular, Visible and Near Infrared: VNIR radiometer data. Moreover, Vegetation Cover: VC, Normalized Difference Vegetation Index: NDVI, Bi-Directional Reflectance

Distribution Function: BRDF of tealeaves have a good correlation to growth index of tealeaves so that it is possible to monitor expected harvest amount and quality of tealeaves with network cameras together with remote sensing satellite data. BRDF monitoring is well known as a method for vegetation growth [3], [4]. On the other hand, degree of polarization of vegetation is attempted to use for vegetation monitoring [5], in particular, Leaf Area Index: LAI together with new tealeaves growth monitoring with BRDF measurements [6].

It is obvious that nitrogen rich tealeaves taste good while fiber rich tealeaves taste bad. Theanine: 2-Amino-4-(ethyl carbamoyl) butyric acid that is highly correlated to nitrogen contents in new tealeaves are changed to catechin [7],[8],[9] due to sun light. In accordance with sunlight, new tealeaves growth up so that there is a most appropriate time for harvest in order to maximize amount and taste of new tealeaves simultaneously.

Depending on the elapsed days after sprouting, tealeaf quality is decreasing. On the other hand, tealeaf yield is increasing with increasing of the days after sprouting. Therefore, there is most appropriate plucking date is very important. Usually, it is determined by the normalized Difference Vegetation Index: NDVI derived from handheld NDVI cameras, drone mounted NDVI cameras, and visible to NIR radiometer onboard satellites because NIR reflection and NDVI depend on tealeaf quality and yield. It, however, does not work well in terms of poor regression performance and species dependency. Moreover, it takes time consumable works for finding appropriate tealeaves from the acquired camera images.

Method for estimation of grow index of tealeaves based on Bi-Directional reflectance function: BRDF measurements with ground-based network cameras is proposed [10]. Wireless sensor network for tea estate monitoring in complementally usage with Earth observation satellite imagery data based on Geographic Information System (GIS) is also proposed [11]. Method for estimation of total nitrogen and fiber contents in tealeaves with ground-based network cameras is, on the other hand, proposed [12].

Monte Carlo ray tracing simulation for bi-directional reflectance distribution function and grow index of tealeaves estimation is conducted with the truth data [13] together with

fractal model-based tea tree and tealeaves model for estimation of well opened tealeaf ratio which is useful to determine tealeaf harvesting timing [14].

Meanwhile, method for tealeaves quality estimation through measurements of degree of polarization, leaf area index, photosynthesis available radiance and normalized difference vegetation index for characterization of tealeaves is proposed [15]. On the other hand, optimum band and band combination for retrieving total nitrogen, water, and fiber in tealeaves through remote sensing based on regressive analysis is discussed [16].

Appropriate tealeaf harvest timing determination based on NIR images of tealeaves is attempted [17] together with appropriate harvest timing determination referring fiber content in tealeaves derived from ground based NIR camera images [18].

Method for vigor diagnosis of tea trees based on nitrogen content in tealeaves relating to NDVI is proposed [19]. In the meantime, cadastral and tea production management system with wireless sensor network, GIS, based system and IoT technology is created [20].

Bi-Directional Reflectance Distribution Function: BRDF model for new tealeaves and tealeaves monitoring with network cameras is well reported [21] together with BRDF model for new tealeaves on old tealeaves and new tealeaves monitoring through BRDF measurement with web cameras [22].

Estimation method for total nitrogen and fiber contents in tealeaves as well as grow index of tealeaves and tea estate monitoring with network cameras is proposed [23]. Meanwhile, multi-layer observation for agricultural (tea and rice) field monitoring is overviewed [24].

The proposed method uses only the days after sprouting. Next thing it has to do is to determination of sprouting date. In order to determine the date, optical sensor onboard Sentinel-2 data is used.

In the following section, the research background is described followed by the proposed method. Then, the experimental method together with experimental results are described. After that, concluding remarks and some discussions are also described.

II. RESEARCH BACKGROUND

Currently, new planting and construction of drink tea gardens are underway in Oita Prefecture, and harvesting has already begun in some of them. As the area for cultivating drink tea grows, it is necessary to improve appropriate management techniques to maintain quality and produce high yields. In this task, we will examine the technique for determining the timely work for growing and picking drink tea.

Setting the first working day is essential for making a picking plan for a large-scale drink corporation. Therefore, it is determined by measuring the Neutral Detergent Fiber: NDF

value (frame picking) and predicting the yield (visually), which is performed after the number of days after sprouting of the field to be picked earliest and after the tea leaves have grown to some extent. The dissemination of technology that allows field managers to easily and accurately predict is an issue.

Growth diagnosis (remote sensing) is performed from image data, and a judgment technique for predicting the plucking time is examined (growth diagnosis based on the correlation between the near-infrared reflectance of tea leaves and the NDF value). However, as will be described later, the tendency differs depending on the variety and changes depending on the weather conditions and the like, so that a very favorable result was not obtained. Therefore, here, we decided to examine a judgment technique for predicting the plucking time based on the correlation between the number of days after sprouting, the NDF value, and total nitrogen.

A. Conventional Method with NDVI

Growth diagnosis (remote sensing) was performed from the image data, and a judgment technique for predicting the plucking time was examined (growth diagnosis based on the correlation between the near-infrared reflectance of tealeaves and the NDF value). The total nitrogen and fiber content of tealeaves are estimated from smartphone camera images acquired by tea farmers, and the optimum growth and plucking time is predicted.

Fig. 1(a) shows the photo of acquiring visible and NIR camera data at the Oita Prefectural Agriculture, Forestry and Fisheries Research Guidance Center: OPAFFRGC. On the other hand, Fig. 1(b) shows an example of the acquired photo of tea trees with visible camera from the top view while Fig. 1(c) shows an example of the acquired photo of tea trees with NIR camera from the top view. In the middle of Fig. 1(c), there is standard plaque which allows conversion from brightness of the images to reflectance. Therefore, it is possible to calculate NIR reflectance of the tealeaves in concern by comparing between brightness of the standard plaque and tealeaves.

From these images and the NDF and Total Nitrogen: TN content in the harvested tealeaves, regression analysis can be done.

B. Species Dependency between NDF and TN and NIR Reflectance

Fig. 2 shows the species dependency between NDF&TN and NIR reflectance. Tealeaf species presented here are Okumidori, Fushun, Sayama-Kaori, Meiryoku, and Yabukita. All varieties showed a positive correlation with NDF and a negative correlation with total nitrogen. The relations between NDF&TN and NIR reflectance, however, are different among the tealeaf species.

Table I shows the results from the linear regressive analysis between NDF&TN and NIR reflectance. The determination coefficients (r^2) are difference each other of the tealeaf species.



(a) Scenery of Acquisition of Photos of tee Trees.

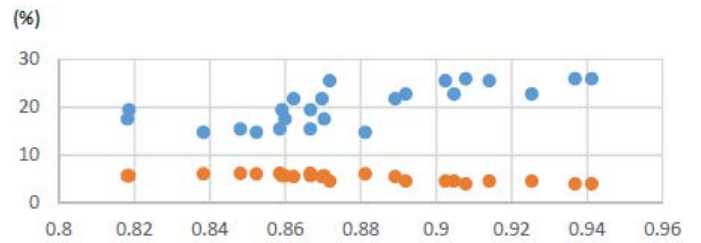


(b) Visible.

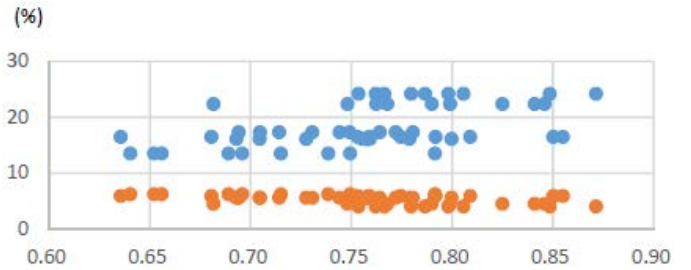


(c) NIR.

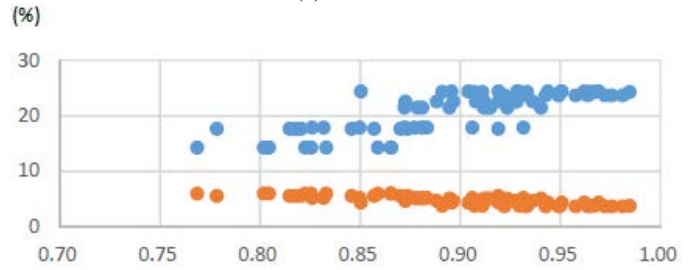
Fig. 1. Scenery of Acquiring Visible and NIR Camera Images and some Examples of the Acquired Photo of Tea Trees with Visible and NIR Cameras from the Top View.



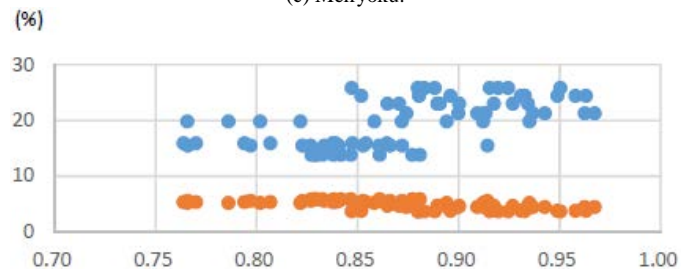
(a) Okumidori.



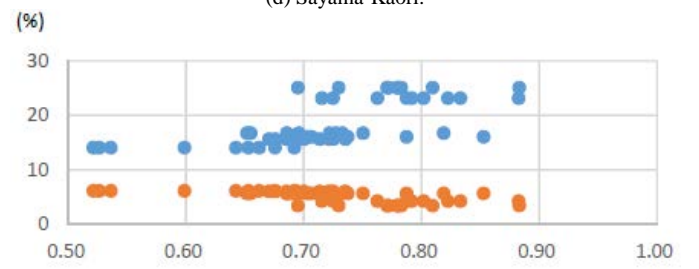
(b) Fushun.



(c) Meiryoku.



(d) Sayama-Kaori.



(e) Yabukita.

Fig. 2. Species Dependency between NDF&TN and NIR Reflectance.

TABLE I. RESULTS FROM THE LINEAR REGRESSIVE ANALYSIS BETWEEN NDF AND TN AND NIR REFLECTANCE

	NDF(%)		TN(%)	
	Linear Approx.	r ²	Linear Approx.	r ²
Okumidori	y=89.97x-58.47	0.52	y=-18.6x+21.62	0.66
Sayama-Kaori	y=50.36x-24.26	0.37	y=-9.27x+12.98	0.38
Fushun	y=34.03x-7.41	0.25	y=-6.75x+10.41	0.23
Meiryoku	y=53.34x-27.04	0.61	y=-11.89x+15.47	0.63
Yabukita	y=34.89x-6.63	0.42	y=-8.09x+11.08	0.39

As a result of correlation analysis and simple regression analysis for each variety of tea leaves with near-infrared reflectance, NDF, and total nitrogen, the near-infrared reflectance increased with the lapse of growing days, and there was a positive correlation with NDF and total nitrogen. There was a negative correlation. In addition, the contribution rate (r²) of near-infrared reflectance varied among varieties for both NDF and total nitrogen, and none showed a strong correlation exceeding 0.7. There was a weak correlation between near-infrared reflectance, NDF, and total nitrogen content, and the results differed depending on the variety.

III. PROPOSED METHOD

The proposed method is based on the days after sprouting. Namely, the most appropriate harvest time can be determined with the days after sprouting. In order to determine the sprouting date, optical sensor onboard Sentinel-2 data is used. Sentinel-2 acquires 10 m resolution of visible to NIR sensor data every 10 days. Therefore, trend of the NIR reflectance can be derived from the sensor data.

Usually, NIR reflectance is increased after the spring pruning (Late March). Then plucking is made in Early May. Within that period, Sentinel-2 derived NIR reflectance can be gathered 5-6 times. Therefore, using these at least three time of acquired NIR reflectance, it is possible to determine the sprouting date which results in determination of the most appropriate plucking and harvest date.

IV. EXPERIMENT

A. Intensive Study Area

The intensive study area is situated at Bungo Ohno in Oita Prefecture, Japan. Fig. 3 shows the location of our intensive study area. There is experimental tea farming area in which several species of tea trees (Okumidori, Fushun, Sayama-Kaori, Meiryoku, and Yabukita) are planted.

B. Estimation of Sprouting Date

Sprouting date can be determined by time series of Sentinel-2 of NIR reflectance, as aforementioned. All the Sentinel-2 of false colored imagery data during from Autumn pruning to just before the plucking are gathered. Fig. 4 shows such imagery data which are covered with no cloud.

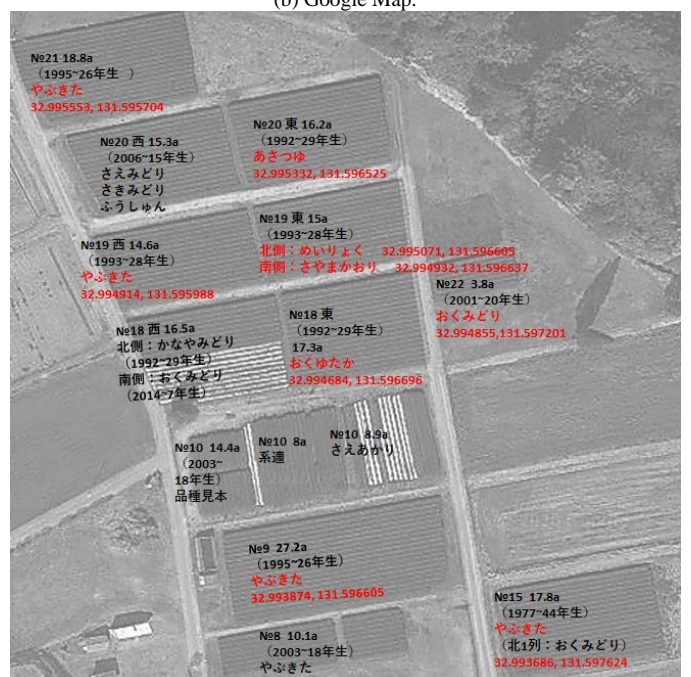
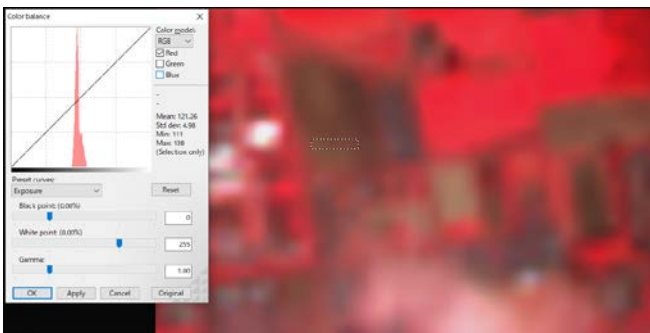
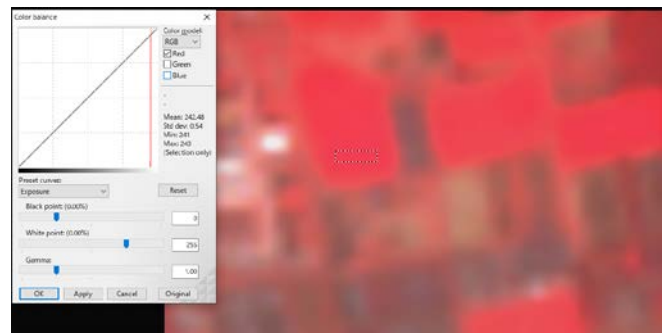


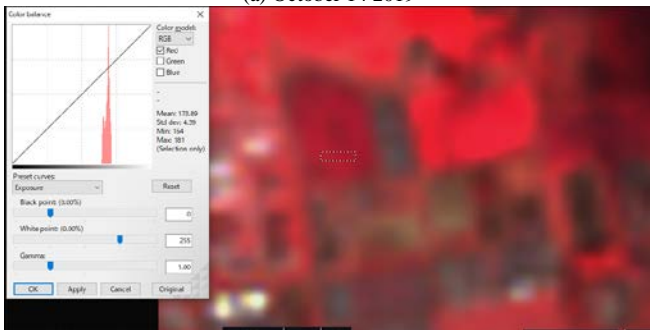
Fig. 3. Intensive Study Area of Bungo Ohno, Oita, Japan.



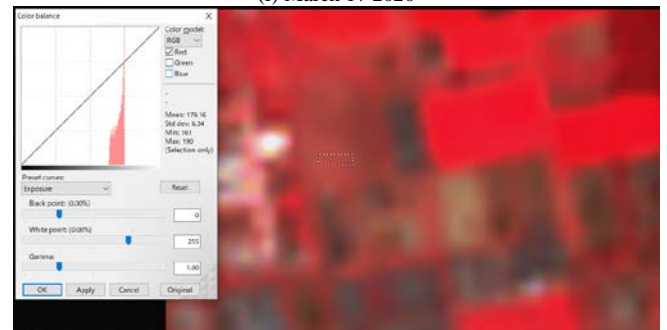
(a) October 14 2019



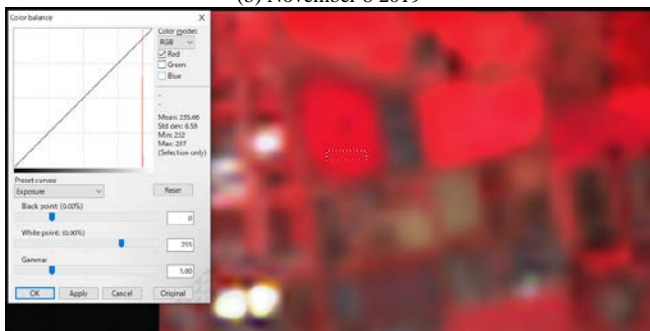
(f) March 17 2020



(b) November 8 2019



(g) April 6 2020



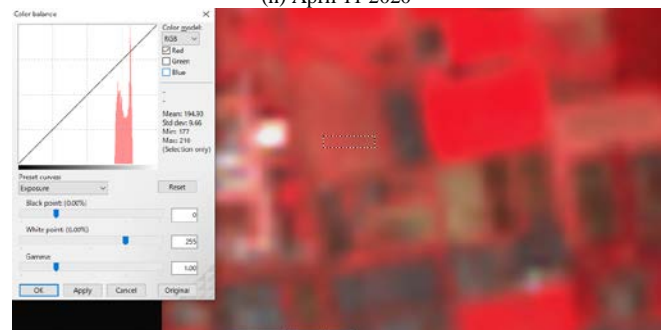
(c) November 13 2019



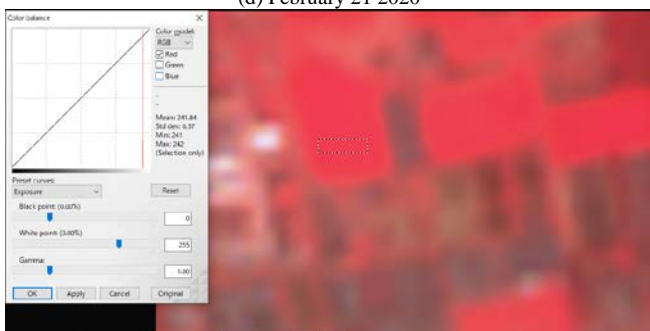
(h) April 11 2020



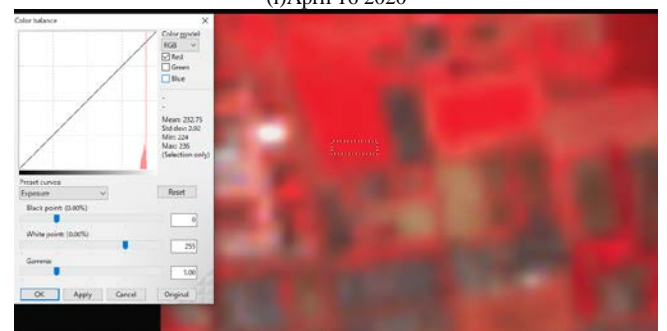
(d) February 21 2020



(i) April 16 2020



(e) March 12 2020



(j) May 1 2020

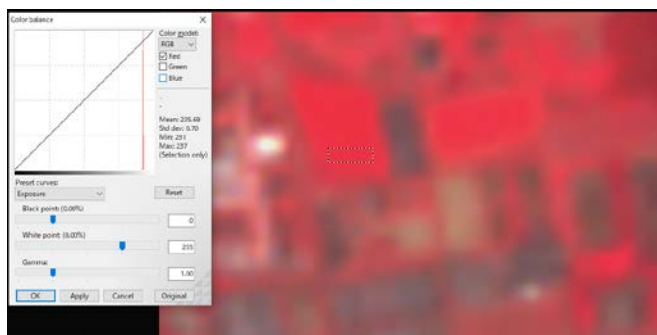


Fig. 4. Sentinel-2 of False Colored Imagery Data during from Autumn Pruning to Just before the Plucking.

In the images, histograms of NIR DN (Digital Number) representing reflectance at the intensive study area are shown. Fig. 5 shows the time series of the NIR DN. Autumn pruning is done on the 14th October 2019. Therefore, tips of tea trees are cut which results in decreasing of NIR DN. After that, tealeaves are growing rapidly. Then NIR DN is saturated during the winter season. During from late March to the bigging of April, spring pruning is done for strength of tealeaves' vitality. Then tealeaves are grown rapidly with new flesh tealeaves. This new flesh tealeaves (Ichiban-Cha) taste good and is contained with Amino acid of Theanine and are to be tealeaves for sale.

From this time series of NIR DN data derived from Sentinel-2 NIR data which are acquired during from Spring pruning to just before the plucking (Fig. 6), sprouting date can be determined. Linear approximation is done with the time series of NIR DN data (four points of data) of Yabukita tea farming field as an example. Then sprouting date is determined as the end of March in this case. The actual sprouting dates for each tea tree species are as follows, Okumidori: 7th April, Fushun: 28th March, Sayama-Kaori: 3rd April, Meiryoku: 27th March, and Yabukita: 30th March. Therefore, the proposed method is validated.

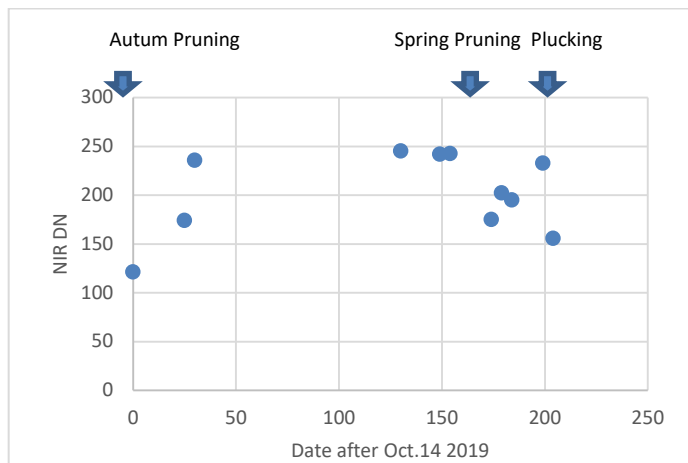


Fig. 5. Time Series of NIR DN Data Derived from Sentinel-2 of NIR Data.

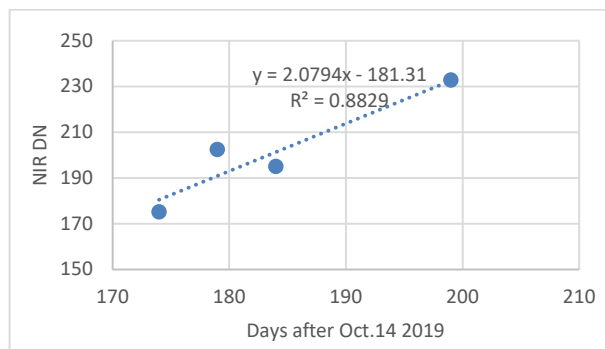
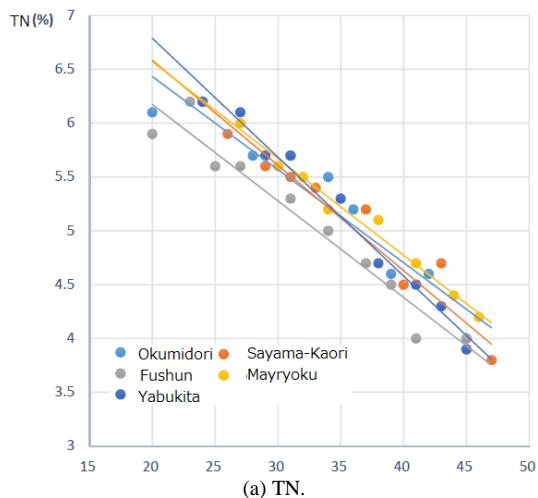


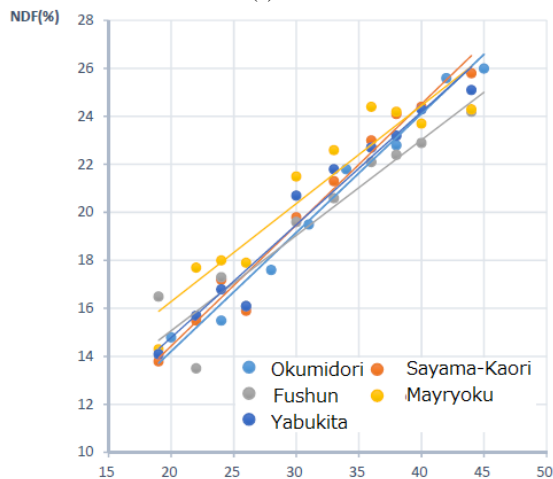
Fig. 6. Time Series of NIR DN Data Derived from Sentinel-2 NIR Data which are acquired during from Spring Pruning to just before the Plucking.

C. Linear Regressive Analysis between TN and NDF and the Days after Sprouting

Linear regressive analysis between TN and NDF and the days after sprouting is done for all the species. Fig. 7 shows scatter plots and linear approximation of the relation between both.



(a) TN.



(b) NDF.

Fig. 7. Result from the Linear Regressive Analysis between TN and NDF and the Days after Sprouting is done for all the Species.

Table II shows the results from the linear regressive analysis between NDF and TN and the days after sprouting. The determination coefficients (r^2) are difference each other of the tealeaf species.

TABLE II. RESULTS FROM THE LINEAR REGRESSIVE ANALYSIS BETWEEN NDF AND TN AND THE DAYS AFTER SPROUTING

	NDF (%)		TN (%)	
	Linear Approx. Eq.	r2	Linear Approx. Eq.	r2
Okumidori	$y=0.4949x+4.3131$	0.98	$y=-0.0868x+8.1638$	0.91
Sayama-Kaori	$y=0.504x+0.3214$	0.97	$y=-0.0947x+8.5326$	0.94
Fushun	$y=0.3975x+6.3236$	0.89	$y=-0.0896x+8.0562$	0.95
Meiryoku	$y=0.4082x+4.4508$	0.89	$y=-0.0926x+8.5441$	0.97
Yabukita	$y=0.4702x+2.5584$	0.96	$y=-0.1121x+9.1617$	0.99

As a result of correlation analysis and simple regression analysis of NDF and total nitrogen with the number of growing days from the germination stage of Ichiban-Cha, the rate of increase of NDF was about 0.4 to 0.5% per day, and the contribution rate of the number of growing days (r^2) showed a strong correlation of 0.89 or higher for all varieties. In addition, the rate of increase in total nitrogen was about -0.09 to -0.11% per day, and the contribution rate (r^2) of the number of growing days showed a strong correlation of 0.91 or more for all varieties.

V. CONCLUSION

Method for most appropriate plucking date determination based on the elapsed days after sprouting with Near Infrared: NIR reflection from Sentinel-2 data is proposed. Depending on the elapsed days after sprouting, tealeaf quality is decreasing. On the other hand, tealeaf yield is increasing with increasing of the days after sprouting. Therefore, there is most appropriate plucking date is very important.

Usually, it is determined by the normalized Difference Vegetation Index: NDVI derived from handheld NDVI cameras, drone mounted NDVI cameras, and visible to NIR radiometer onboard satellites because NIR reflection and NDVI depend on tealeaf quality and yield. It, however, does not work well in terms of poor regression performance and species dependency. Moreover, it takes time consumable works for finding appropriate tealeaves from the acquired camera images.

The proposed method uses only the days after sprouting. Next thing it has to do is to determination of sprouting date. In order to determine the date, optical sensor onboard Sentinel-2 data is used. Through experiment with the truth data taken at the intensive study area of the Oita Prefectural Agriculture, Forestry and Fisheries Research Guidance Center: OPAFFRGC, it is found that the proposed method is validated.

VI. FUTURE RESEARCH WORKS

Further experimental studies are required for further validation of the proposed method for determination of the most appropriate plucking date for harvesting good quality of tealeaves of new flesh tealeaves (Ichiban-Cha).

ACKNOWLEDGMENT

The author would like to thank Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

REFERENCES

- [1] J.T.Compton, Red and photographic infrared linear combinations for monitoring vegetation, *Journal of Remote Sensing of Environment*, 8, 127-150, 1979.
- [2] C.Wiegand, M.Shibayama, and Y.Yamagata, Spectral observation for estimating the growth and yield of rice, *Journal of Crop Science*, 58, 4, 673-683, 1989.
- [3] S.Tsuchida, I.Sato, and S.Okada, BRDF measurement system for spatially unstable land surface-The measurement using spectro-radiometer and digital camera- *Journal of Remote Sensing*, 19, 4, 49-59, 1999.
- [4] Kohei.Arai, Lecture Note on Remote Sensing, Morikita-shuppan Co., Ltd., 2000.
- [5] Kohei.Arai and Y.Nishimura, Degree of polarization model for leaves and discrimination between pea and rice types of leaves for estimation of leaf area index, Abstract, COSPAR 2008, A3.10010-08#991, 2008.
- [6] Kohei.Arai and Long Lili, BRDF model for new tealeaves and new tealeaves monitoring through BRDF monitoring with web cameras, Abstract, COSPAR 2008, A3.10008-08#992, 2008.
- [7] Greivenkamp, John E., *Field Guide to Geometrical Optics*. SPIE Field Guides vol. FG01. SPIE. ISBN 0-8194-5294-7, 2004.
- [8] Seto R H, Nakamura, F. Nanjo, Y. Hara, *Bioscience, Biotechnology, and Biochemistry*, Vol.61 issue9 1434-1439 1997.
- [9] Sano M, Suzuki M ,Miyase T, Yoshino K, Maeda-Yamamoto, M.,*J.Agric.Food Chem.*, 47 (5), 1906-1910 1999.
- [10] Kohei Arai, Method for estimation of grow index of tealeaves based on Bi-Directional reflectance function: BRDF measurements with ground-based network cameras, *International Journal of Applied Science*, 2, 2, 52-62, 2011.
- [11] Kohei Arai, Wireless sensor network for tea estate monitoring in complementally usage with Earth observation satellite imagery data based on Geographic Information System (GIS), *International Journal of Ubiquitous Computing*, 1, 2, 12-21, 2011.
- [12] Kohei Arai, Method for estimation of total nitrogen and fiber contents in tealeaves with ground-based network cameras, *International Journal of Applied Science*, 2, 2, 21-30, 2011.
- [13] Kohei Arai, Monte Carlo ray tracing simulation for bi-directional reflectance distribution function and grow index of tealeaves estimation, *International Journal of Research and Reviews on Computer Science*, 2, 6, 1313-1318, 2011.
- [14] Kohei Arai, Fractal model-based tea tree and tealeaves model for estimation of well opened tealeaf ratio which is useful to determine tealeaf harvesting timing, *International Journal of Research and Review on Computer Science*, 3, 3, 1628-1632, 2012.
- [15] Kohei Arai, Method for tealeaves quality estimation through measurements of degree of polarization, leaf area index, photosynthesis available radiance and normalized difference vegetation index for characterization of tealeaves, *International Journal of Advanced Research in Artificial Intelligence*, 2, 11, 17-24, 2013.
- [16] Kohei Arai, Optimum band and band combination for retrieving total nitrogen, water, and fiber in tealeaves through remote sensing based on regressive analysis, *International Journal of Advanced Research in Artificial Intelligence*, 3, 3, 20-24, 2014.
- [17] Kohei Arai, Yoshihiko Sasaki, Shihomi Kasuya, Hideto Matsuura, Appropriate tealeaf harvest timing determination based on NIR images of tealeaves, *International Journal of Information Technology and Computer Science*, 7, 7, 1-7, 2015.
- [18] Kohei Arai, Yoshihiko Sasaki, Shihomi Kasuya, Hideo Matsuura, Appropriate harvest timing determination referring fiber content in tealeaves derived from ground based NIR camera images, *International Journal of Advanced Research on Artificial Intelligence*, 4, 8, 26-33, 2015.

AUTHORS' PROFILE

- [19] Kohei Arai, Method for Vigor Diagnosis of Tea Trees Based on Nitrogen Content in Tealeaves Relating to NDVI, International Journal of Advanced Research on Artificial Intelligence, 5, 10, 24-30, 2016.
- [20] Kohei Arai, Cadastral and Tea Production Management System with Wireless Sensor Network, GIS, Based System and IoT Technology, International Journal of Advanced Computer Science and Applications, 9, 1, 38-42, 2018.
- [21] Kohei Arai, Lili, Long --BRDF model for new tealeaves and tealeaves monitoring with network cameras, Saga University Faculty of Science and Engineering Bulletin, 38, 1, 23-28, 2009.
- [22] Kohei Arai and Long Lili, BRDF model for new tealeaves on old tealeaves and new tealeaves monitoring through B RDF measurement with web cameras, Abstract of the 50th COSPAR(Committee on Space Research/ICSU) Congress, A3.1-0008-08 ,992, Montreal, Canada, July, 2008.
- [23] Kohei Arai, Estimation method for total nitrogen and fiber contents in tealeaves as well as grow index of tealeaves and tea estate monitoring with network cameras, Proceedings of the IEEE Computer Society, Information Technology in Next Generation, ITNG, 595-600, 2009.
- [24] Kohei Arai, Multi-Layer Observation for Agricultural (Tea and Rice) Field Monitoring, Proceedings of the Seminar at Bogor Agriculture University, Keynote Speech, 2016.

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 Mister 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>.

Systems Security Affection with the Implementation of Quantum Computing

Advances in Quantum Computing

Norberto Novoa Torres¹, Juan Carlos Suarez Garcia², Erik Alexis Valderrama Guancha³

Faculty of Technology, Telematics Engineering
Universidad Distrital Francisco José de Caldas
Bogotá, Colombia

Abstract—Current security systems use cryptographic robust tools that have been of great help in regulating information. During its time the implementation of these tools abolished the classic security systems, as by means of cryptanalysis they allowed decryption of information in a fast, automated, and simple mode from these systems. Considering this scenario, the same happens when quantum cryptographic systems are implemented, insomuch as the current security systems could be abolished, as tools exist that permit its encryption in a simple way, but with the risk of putting the data of worldwide organizations in danger. With the purpose of mitigating these risks, it is necessary to consider the upgrade of the available security systems, by security systems and quantic encryption, before a massive implementation of the quantum computer's use as an everyday tool. With this it does not mean that quantum computing would be a disadvantage, on the contrary, the advantages from this technology will mean that security information and data are almost invulnerable, which is a meaningful advance in the IT field. With security information professionals are obliged to recommend and perform an appropriate migration of new technologies to avoid existing exposition risks as data as well as transactions. If this were not the case, the same scenario presented in the classic security systems would occur.

Keywords—Quantum computing; encryption; cryptography; cryptanalysis; data security

I. INTRODUCTION

Encryption systems are drawn from the Vth century BC with the emergence of the Spartan Scythian, which consists of two bars of the same width in which people could only read messages. Then Caesar's encryption system appeared which is based on mono-alphabetic substitution. Another important encryption system is the Vigenere that implements the multi-alphabetic with character matrices. These systems are the ones we know today as classic encryption systems. During the Second World War, in the XX century, calculating machines were implemented with the purpose of deciphering the messages from these encrypted systems. Here the Turing machine is highlighted whose principle was based on rotors that automate the calculations to decrease the encryption time of the classic cipher methods, resulting in the obsolescence of these methods and the search for more complex encryption implementation methods. At this time, due to computation speed and the then current computers, new encryption methods

were implemented, as the ones based on private password cryptography, in which is used the same password to cipher as well as decipher, are known as symmetric cryptography methods. Among them we could mention the DES algorithm that divides the messages into two similar parts and each interaction works alternately implementing an initial and final permutation. Other encryption methods are based on public password cryptography known as asymmetric key cryptography methods. These use a pair of passwords that belong to the same person, one is public and the other is private, for instance the Diffie-Helman and RSA algorithm. We can say that current cryptography is divided into two main types, symmetric key cryptography and asymmetric key cryptography. With the implementation of these current cryptographic systems, we can reflect about what happened to the classic encryption systems which were made obsolete due to emergence of current cipher methods. The classic methods could be decrypted in a much shorter time. Subsequently, we can see that the advance in quantum computing investigation will contribute to advancing current computing as we know it. This is based on the use of Qubits, a combination of ones and zeros, instead of bits as in the computers used currently., Also cipher algorithms would be on this principle, like the (QKD) quantum key distribution algorithm, Peter Shor, Groover and McElice Algorithm. These quantum encrypted systems are more robust and capable of ciphering and deciphering information in such a way that they are almost impossible to break. From the perspective of the symmetric and asymmetric key encrypted systems used recently, their security is based on the use of complex mathematical problems which would take years to solve. However, with the implementation of quantum computers, these mathematical problems will change from being solved in years to minutes, as their processing speed will be much higher, and additionally will put in check security safeguarded information.

II. METHODOLOGY

A. Cryptography

Cryptography is the creation of techniques to encrypt data, having as its objective to obtain the confidentiality of the messages. If cryptography is the creation of mechanisms to cipher data, the cryptanalysis is the method to "break" these mechanisms and obtain the information. Once our data has passed a cryptographic process, we would say the information

is encrypted. Cryptography is a word that comes from Greek “Kryptos” that means hidden, whereby, it is understood as the study of science which by the treatment of information, protects itself from modifications and unauthorized use. Cryptography, besides being a discipline that studies the principles, methods and means to transform data to hide its meaning, guarantees its integrity, establishes its authenticity, and prevents its rejection. This has current mathematics bases that are: number theory, algorithmic complexity theory, information and statics theory. Cryptography must ensure that the sent information is authentic in a double sense: the sender is really who it says it is and that the content of the sent message, normally called the cryptogram, has not been modified during the transit. According to the Greek historian Polibio, the first cryptosystem was a substitution system based on the position of a letter in a table remarkably like Caesar’s system used by the Romans, for example in military campaigns. Another documented cryptosystem that existed was the Scythian Spartan one which was based on a method of transcription using a cylinder as a key rolled to cypher and decipher.

Starting from the Second World War in the XX century, cryptography used tools like calculation machines to be more robust. The best known is the German Enigma machine, that used rotors considerably automatizing the calculations. This was necessary to perform the cipher and decipher operations from the messages that later were decoded by the mathematician Alan Turing. The classic methods use substitution and transposition methods over the characters in a message [1], these techniques having been proposed by Shannon to accomplish confusion and diffusion:

- Reverse Transposition: Consists in inversion of the message from the beginning to the end.
- Simple Transposition: This method divides the message symbol by symbol. If the total of symbols is odd for this group, one more symbol is added, getting two groups, the first one odd and the second one paired. They are united and the encrypted messaged is obtained.
- Double Transposition: The simple transposition is applied twice.
- Transposition by Groups: It is traded in a way the text characters are reordered in n character blocks, but they are reordered with a position number in the cryptogram, for example 321. meaning that character 3 is transmitted first, then the second, finally the first. This is repeated continuously by 3-character blocks. To decipher the message in this case the password would be 321.
- Transposition by Series: The message is ordered in such a way that the cryptogram is made by the sequence of the messages that has been considered to create it. In other words, simple functions are presented in a specific order to be able to cipher and decipher the message.

Modern cryptography is divided into two main tracks: key symmetric and key asymmetric cryptography. Symmetric cryptography, or secret key, are those algorithms that use only one cipher and decipher key. Therefore, its diffusion must be protected, as it is only for the authorized sender/receiver. For instance, the Data Encryption Standard algorithm or (DES) “Fig. 1” created in the 70s by IBM [2], uses the Feister framework (The data blocks are divided into two equal parts and in each interaction each one of its parts works alternatively) in blocks of 64 bits. Its initial key is 64 bits, then by each interaction it generates one of 56 bits, altogether it works with 16 interactions, implementing an initial and final permutation [3].

By contrast the asymmetric key cryptography (also known as public key one) is a system that implements a pair of keys. This pair of keys belongs to the same person. One is in the public domain, the other one is private. The MD5 algorithm “Message Digest Algorithm” is a coding algorithm related to a file. The one used to verify the file itself has not been modified “Fig. 2”, it was designed by Ronald Rivest in 1991 [4]. It is based on a cryptographic reduction of 128 bits and is 5 parted: 1) Bits Addition: taking from an original text, this text is extended until it is consistent with the number 448 and module 512, to this is added a bit “1” then bit zero “0” to be extended 2) Length Message: a whole number of 64 bits is calculated which is the original text length before step 1 is made. These bits are linked to the result in step 1 having as a result a length that is multiple of 512. 3) Start a MD buffer: It is a 4-word buffer where each one has a length of 32 bits, and they are used to calculate a summary of the text. 4) Processing the text: XOR, AND, OR and NOT types of operations are made, also using a 64 elements table, made from a sine function, resulting in 4 words of 32 bits. 5) Exit: Finally, an exit text is produced where the 4 words come out from the least weight to the greatest.

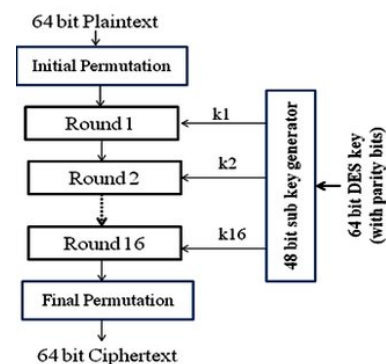


Fig. 1. DES Algorithm.

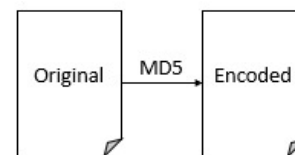


Fig. 2. MD5 Algorithm.

B. Quantum Cryptography

Although, traditional cryptography nowadays allows keeping a secure communication between two parts, the algorithmic proposals of Shor in 1994, which use the characteristics of a hypothetical quantum computer, could put in danger some of the most used cryptographic systems, such as the RSA. Quantum cryptography can reach secure communications using natural laws in a quantum scale, such as Heisenberg's Uncertainty Principle, of quantum overlaying and quantum interlinking. Quantum mechanics is a model to describe the behavior of subatomic particles [5]. With this it is proven that reality at a microscopic level behaves very differently to that at a macroscopic one. Classic physics, particularly Newton's Laws of Mechanics allow the performance of experiments that are verified at a macroscopic scale. For example, it is possible to track the journey of a cannon ball, or a basketball. However, quantum physics occurs at a level that is impossible to perceive with the five senses, like the polarization of a photon, or the spin of an electron.

Classic computers are physical systems, i.e., implications in terms of space, time and energy. Over time the demand for faster computers is higher. While computation devices keep miniaturizing, they get closer to the microscopic level, in which the laws of the quantum world rule. By the year 2020, computation will be carried on at an atomic level. With the arrival of quantum theory and some of its characteristics, like quantum overlaying and quantum interlinking, it was predicted that quantum computers, defined as "a type of computer that explodes the interactions of quantum mechanics", could develop some computing tasks exponentially faster than any conventional computer. These predictions go with the development of quantum algorithms, which from its theory, take advantage of quantum mechanics' features. To factor a number with 400 digits, a numerical achievement needed to break some security codes, it will take billions of years with even the fastest supercomputer. However, a quantum computer could complete the task in about one year.

C. Quantum Cryptographic Models

Quantum computing has created cryptographic models that work under the same paradigm, to achieve sharing information in two parts safely. Nevertheless, its construction and definition depend more on physics than on mathematics, because for this time the passwords are transmitted in photon shapes, not in bits as the current cryptography [6]. To understand these quantum cryptographic models, it is necessary to understand the following concepts:

- **Superposition Principle:** This principle consists in a quantum particle that is in a state of superposition. This means, it behaves as having different states at the same time [7], to change the state of the spin, the energy unit must be used, for example, laser, but what could happen if only half of the energy is provided? In this case theory says that the particle turns into a state of superposition.
- **Heisenberg Uncertainty Principle:** This principle establishes that in the subatomic world is not possible to know the values of two different magnitudes, from an

elemental particle at the same time. Because the fact is that measuring the first one interferes with our capacity to measure the second one.

- **The Shannon Bit:** This consists in a bit only which can take one value at the same time, it can be 0 or 1. These bits have the ability to be copied.
- **Qubit:** This is that quantum unity which can take different values at the same time. It does mean 0 and 1 at the same time, so different from the Shannon bit.
- **No-cloning Theorem:** This theorem is the result from the quantum mechanics that bans the creation of identical copies from unknown and arbitrary quantum states.
- **Quantum Entanglement:** Starting from a group of particles that are entangled or connected in their existence, which even though they are separated by thousands of light years, a change of state of one of them affects the rest immediately.

D. Quantum Key Distribution (QKD)

The algorithm of Quantum Key Distribution (QKD) is a revolutionary encrypted technology that takes advantage of quantum mechanics' laws to reach a secure key interchange in the information theory. QKD allows the two parts of the encryption process "to increase" a secret shared key without adding any limits in the power of computing processing and is one of a kind in its capacity to detect the presence of any third party's participation during the keys' interchange. Due to the fundamental laws of quantum mechanics, any interception from a third party during the key interchange will introduce traceable errors. If the errors are under a defined limit, a secure unconditional key can be distilled. When a QKD is used with a symmetric cryptographic algorithm like the One-Time Pad [8], the result is an unconditionally cryptographic secure system. The beginning of the quantum keys distribution (QKD) dates to Stephen Wiesner, who developed the idea of quantum codification of concerted in the last decade of 1960. He described two applications of quantum codification: a fraud proof creation money bills method (quantum money), also a method for the creation of multiple messages in such a way that one reading of the messages destroys the others (quantum multiplexing). The Wiesner quantum multiplexing uses polarized photons in concerted bases like quantum bits (qubits) to pass the information. Thus, if the receiver measures the photons in the correct polarization base, it will get a high probability of a correct result. However, if the receiver measures the photons on the incorrect base, it will obtain a random result and it will destroy all the information on the original base. The encrypted process is the following, to destroy a key, the dispenser part, Alice, creates a random bit in a random base, sending a photon this being 1 or 0 in a horizontal or vertical way. The photon created by Alice is received by Bob, who does not know the base Alice used. Bob measures the photon polarization, as he chooses randomly any of the two bases. Alice and Bob then discuss through a public channel on the one they decide to measure, and they discard the bits Bob did not measure using the same base that Alice did to create the photons.

This process permits a secure channel to the key distribution, as any person that listens to the channel must guess in which base it measures. If Alice and Bob choose the same base, but the spy chooses a different base, there is a 50% possibility that Bob will measure a different bit value from the one Alice sent. Therefore, Alice and Bob have the possibility to detect an interceptor through a public comparison and discard a certain number of bits for the ones who chose the same base.

E. Peter Shor Algorithm

The Peter Shor Algorithm allows decomposing prime factors in any N number, hence its potential implementation in a quantum computational device brings as a consequence that cryptographic systems based on a factorization process like the public key system RSA will be broken easily [9]. While the processes related to the key public algorithms are executed in super polynomial times by the mode $\exp[c(\ln N)^{1/3} (\ln)^{2/3}]$, to the Shor quantum algorithm, with the necessary time to execute this same polynomial task and by the mode $O(\log(N)^3)$. The great strength of calculation the Shor algorithm has regarding the implemented algorithms in conventional computer consists in the fact of doing quantum effects like interference, enhancing and allowing an information process in a parallel mode, which is competingly translated in a processing reduction of time. This algorithm underlies its power to determine the period of an adequate function. Although its study presents a high level of complexity, it can be interesting to analyze the new approach of quantum mechanics that offers a solution to the problem of factorization.

F. Groover Algorithm

The Groover Algorithm is used to search in a non-ordered sequence of data. It was invented by the North American scientist of Indian origin, Loc Kumar Grover in 1996 [10]. This algorithm avoids reconstructing a previous organization of the search. The algorithm is strictly probabilistic whose answer has an error percentage and so must be small enough. To explain how it works let us suppose we have a non-structured data base with N elements, and they are numbered from 0 to N-1. These elements are not in order. Normally, we will test element by element, until we have the one, we are looking for. Using the Grover Algorithm, *we only need attempts*. The Groover algorithm has registrations: n qubits in the first and 1 qubit in the second. The first step is to create a superposition of all 2n states from the computing base. This is made by initializing the first register in the state and applying the operator H_n , where H is from the Hadamard door [11]. Then, we set a function f that recognizes the solution as $f: \{0, \dots, N-1\} \rightarrow \{0, 1\}$, $f(k) = 1$ if K is the searched element, on the contrary $f = 0$. The function f “Fig. 3” also is recognized as an oracle and it can be defined as:

This algorithm has an operator sequence of Groover (G) iteration and the states of the first register correspond to the first iteration.

G. McEliece Encryption

The McEliece cryptosystem is an asymmetric encryption algorithm developed in 1978 by Robert McEliece. It was the first of these schemes to use randomization with the encryption

process. The algorithm has never earned much acceptance among the cryptographic community, but it is an application of the quantum post-cryptographic, as it is immune to attacks using the Shor algorithm and, more generally, the measure of the coset states using the Fourier sampling.

The algorithm is based on the strength of decoding a general linear code [12]. To a decryption of the private key, a correction code of errors is selected as it is known that a decoding efficient algorithm can correct the errors. The original algorithm uses binary Goppa codes (codes subfield of geometrics, Goppa codes from a gender-0 curve over finite fields of characteristics. These codes are easy to decrypt due to an efficient algorithm on account of Patterson. The public key drifts from the private key to encrypt a selected code as a general linear code “Fig. 4”. The McEliece cryptosystem has some advantages, for example, with RSA the cipher and decipher are faster. For some time, it was thought McEliece could not be used to generate signatures. However, a signature scheme can be built over the Neiderreiter base scheme, the double alternative of the McEliece scheme. One of the main disadvantages of McEliece is that both private and public keys are large matrices. To a standard parameter’s selection, the public key is 512 kilobits. Because of this, the algorithm is used too little during practice. An exceptional case that uses McEliece for encryption is the Freenet application. McEliece is about three algorithms: a probabilistic key generation algorithm that produces a public and private key, a probabilities algorithm, and a deterministic decryption algorithm.

H. Quantum Cryptography vs. Classic Cryptography

As it is seen during this paper, one can analyze how the techniques, modes, and methods to make a cryptographic system securer have evolved, from simple steps or movements with the alphabet to complex mathematical calculations, for instance, one technique is the use of substitutions. This consists in replacing by parts the message to transmit with other words or symbols keys that later can be decrypted by the receiver if this one has the same key combination. Another one is by making a transposition in the message in which one will not be replaced by other alphabet, but they change them in position or geometrical mode order, for example to write the message backwards and at the end by making mathematical calculations with prime numbers, modulation of a number, whole factorization and discrete logarithms, among others.

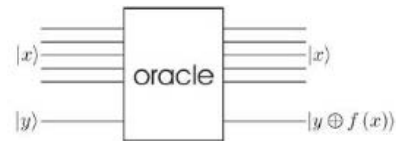


Fig. 3. Oracle of the Function.

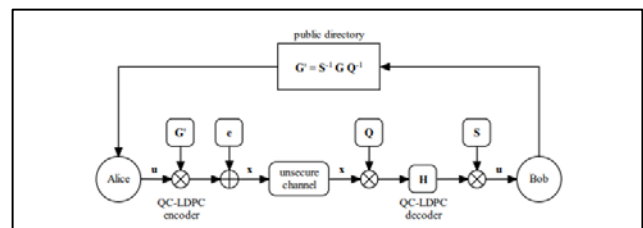


Fig. 4. McEliece Encryption.

These techniques have solved the problem of transmitting messages in a secure mode through time, in most telecommunications such as cellphones, the internet, fixed telephony, satellites, the use of cryptography exists. Similarly, in the financial field when the use of cryptography was adopted by banks, this necessitated the change in use of cards with magnetic stripes to the use of EMV chip cards, with the calculation capacity, and the holders' reliability, improving the quality and security of transactions. Large financial companies like Visa, Mastercard, American Express adopt this technology and create their own security standards [13].

Due to cryptography our data has managed to travel in different networks safely, however, there are also those who want hidden information and with the use of cryptographic engineering have managed to decipher algorithms that were believed impossible to break. Every time the complexity of a cryptographic algorithm increases, more processing memory is needed. A common electronic device such as the cellphone does not have the capacity to process a complex algorithm. The human mind can create new algorithms to hide information, whereas the machine is still inferior in its processes. There are mathematical algorithms which the traditional machine can take years to process and solve, such as the Riemann Hypothesis where all non-trivial zeros of the Riemann zeta function have a real part of $1/2$. These complex mathematical processes are processed in large supercomputers that are not available for anyone as they use a lot of space, consume a lot of energy, and are expensive. Thus, it is impossible for a device as small as a cellphone to encrypt a message that is totally safe for several years. It should be noted that existing algorithms are still being improved to make them more secure. But it cannot be forgotten that there are always the same bases and principles to hiding a message, making it vulnerable, because, just as the digital age advances rapidly, breaking security systems also advances. This is classical cryptography, which is accustomed to employing the same bases, the same tools, and the same laws of physics.

The birth of quantum cryptography occurred when several physicists, including Albert Einstein, discovered that there were different and unexpected behaviors when one enters the subatomic world of matter [14], called quantum physics, where many of the physical laws known nowadays do not apply. Normally an algorithm returns something that can be true or false if its execution were correct or not if the light bulb were on or off. When we talk about quantum physics, creating an algorithm at that level is expecting different simultaneous responses which are impossible to be observed at that moment. If one wants to know a precise answer statistical predictions should be used. In classical cryptography if the encryption algorithm is known, a spy can at some point in the execution of the algorithm, find the hidden message. While in quantum cryptography it is impossible to stop at an execution point and find the hidden message since the spy would find a message that is no longer part of quantum physics. That means it would no longer be part of the quantum algorithm, therefore the message is modified making it impossible to reach the receiver. Consequently, this is a point in favor in quantum cryptography because if the receiver at no time receives the hidden message, it is because the channel has been intercepted, while in

classical cryptography the intruder can even listen to the channel without the sender or receiver suspecting it.

Regarding this speed, whereas in classical cryptography when it makes the algorithm more complex the processing time is slower. In quantum physics an algorithm of considerable complexity can be processed in less time. This means that with the use of the quantum physics it can be implemented without any problem ciphered algorithms of high complexity, allowing an increase of the security in communications, including finance, the results are shown in Table I. However, this science is new in practice and the attempts to reach quantum behaviors are highly expensive, as they involve subatomic particles where the matter overheats and its refrigeration is not at all easy. This is a negative point because there still does not exist sufficient portable hardware for computer systems to take advantage of the benefits of quantum mechanics. Even so, at present all the potential that could be had if a quantum computer were created is being studied by means of theories and the creation of new quantum algorithms that can even easily break the existing cryptographic algorithms. This is the case of the Shor Algorithm which is implemented in a quantum computer that can break down into prime factors any number no matter its length. Or $((\log N)^3)$, which could leave obsolete a lot of public key cryptographies like the RSA. An advantage that quantum physics has is entanglement, consisting of, a particle influencing another particle despite how far we find it from the other one. This can yield an advantage for secure communication because while a sender has in its power a particle that influences another particle that the receiver has, it be able to communicate only by taking advantage of this property. Now imagine that it were not only a particle but the whole message, then it would only be required to interlink the messages. Thus, the two points can communicate by means of a quantum channel. This is different from classical cryptography where the sender and receiver must share a key or password that allows enciphering and deciphering of the message. Here below, is shown a comparison between these two types of cryptographies, noting that the quantum is superior.

TABLE I. COMPARISON OF QUANTUM CRYPTOGRAPHY AND CLASSICAL

	Classical	Quantum
Speed	As more complex, slower	Fast, no matter the complexity
Security	Only for a while until deciphering is done	When the message is intercepted immediately it loses the information.
Implementation	For more complex calculations it requires more power of processing.	The hardware of quantum technology is in the development process
Future	Insecure	Decrypts most of the implementations of classical cryptography.
Channel	There must exist a physical connection between the parts	It is not necessary to have a physical connection if it takes advantage of quantum interlinking
Keys	The sender and receiver must have the key to encipher and decipher the message.	With the use of quantum interlinking the use of keys would not be necessary.

I. Types of Security Systems Would be Affected

Future general use of quantum computers would take away the latent risk of the current public key infrastructure systems and private key. The majority of world-wide transactions are based on the systems of current cryptography to protect their security, the super powered quantum computers making use of the encrypted and decrypted quantum algorithms mentioned in the present article, which are based in quantum bits or qubits. These will allow the speed of processing from encryption and decryption of the best systems of security used today to change from being decrypted in thousands of years to only fractions of a second.

Quantum computing is really closer than we imagine. The companies and organizations that handle critical data like banks, governmental agencies and the field of medicine development that do not start considering this scenario, could be exposing data and information highly vulnerable as their methods of encryption could be decrypted easily when facing quantum computing.

If information security professionals do not implement measures right now and understand that the advance in the construction of operable quantum computers is developing in great steps there potentially will be a threat for the current methods of encryption and they will be faced by an exhibition of data and catastrophic leaked information.

The systems of asymmetric encryption would be the most vulnerable as they are based in cryptography of public key, the one that is based on extraordinarily complex mathematical algorithms [15]. However, this prospect would remain delayed if they used quantum computers.

Because of this it may be too late to guarantee that when quantum computation is implemented the processes of encryption in the different organizations are protected because it would take too long for the computation of trail algorithms and quantum encryption to integrate to the organizations satisfactorily. Although there are tasks that security professionals can implement so they are not totally exposed, these are based on encryption agility and the capacity to implement and adapt algorithms of quantum encryption once available.

J. Encrypted Simulation

A quantum computer will break the encrypted RSA of 2048 bits in eight hours. If now a quantum machine of 20 million cubits were implemented it could do it in a record time [16]. The systems of encryption like the RSA never have been unailing. As mentioned, its security is based in the massive quantity of time that a conventional computer would need to do it. The current methods of encryption are designed specifically for encrypted processes that were so slow that they seemed practically unbreakable.

Several computer scientists have tried to calculate the resources that a quantum computer would need to discover how long it would take to build a machine of this type. However, this calculation must be reviewed due to the work of the Google researcher in Santa Barbara (USA) Craig Gidney and the researcher of the Royal Institute of Technology KTH in Stockholm (Sweden) Martin Ekerå. Both have found a more

efficient way for quantum computers to perform encryption code calculations, which reduces the resources needed by a magnitude of several orders.

Their findings suggest that these machines are much closer to being made for real than we had suspected. The effect will be uncomfortable for governments, military and security organizations, banks and anyone who needs to store data for longer periods than 25 years from now, because as is indicated in the algorithm of Peter Shor, big numbers are factorized, and there is the crucial element of the process to decrypt the codes based in mathematic algorithms.

These algorithms are based on the process of multiplication, that is easy to perform in one direction, but much more difficult in the reverse sense. For example, multiply two numbers result very simple: $593 \times 829 = 491.597$. The difficult part is to begin with the number 491.597 and calculate which two numbers have been multiplied to produce it.

As the numbers increase, the issue gets more complicated. In fact, computer scientists consider it practically impossible for a conventional computer to calculate the numbers with more than 2048 bits, which is the base of the most used mode of the enciphered RSA.

Shor showed that a quantum computer powerful enough could do it with ease, something that surprised the cyber security industry. Since then, the power of the quantum machines has not done anything but increase. In 2012, some physicists used a quantum computer of four cubits to calculate the factor 143. In 2014, they used a similar device to calculate the factor 56,153.

So, anyone could think that, at this pace, quantum computers are about to surpass the best conventional computers.

For Gidney and Ekerå have just shown that a quantum computer could do the calculation with only 20 million cubits. In fact, they claim that it would take only eight hours in completing the calculation. As a result, the estimate of the worst case of how many cubits needed to factorize the RSA of 2048 bits has reduced by almost two orders of magnitude. Its method focuses on a more efficient mode to make a mathematical process called modulate exponentiation "Fig. 5". It consists in finding the remainder when a number elevates to a true level and afterwards divides it by another number. This process is the most expensive operation of the computing level of the algorithm of Shor. However, Gidney and Ekerå have found several forms to optimize it, which significantly reduce the necessary resources to execute the algorithm.

For any ordinary citizen, this finding does not entail a lot of risks. Most people use encryption of 2048 bits, or similar, for tasks like sending details of credit cards through the internet. If these transactions were registered today and be decrypted in 25 years, the damage would be minimal. Nevertheless, for governments, the situation would result in more trouble. Today's messages sent between embassies or the army could be important in the next 20 years and would be better kept secret. If these messages were still sent through an enciphered RSA of 2048 bits, or something similar, these organizations would have to begin to worry a lot.

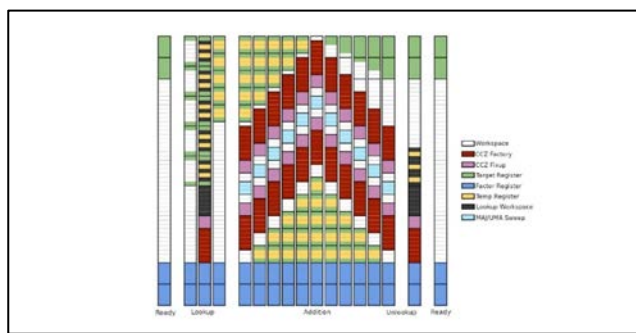


Fig. 5. Calculation Simulator of Encryption.

K. How to Find the Security in the Cloud

Different internet platforms offer cloud services where the user does not require installing any type of software in their computer, rather only using a browser web that can access different services of the cloud, as it is with emails, music, editing of photos and videos, online office, online films, social networks etc. Each service in the cloud has implemented its security based on the protocols of SSL (Secure Sockets Layer) and TLS (Transport Layer Security) These are cryptographic systems for computer networks that guarantee transmitted information in the network is not intercepted by an unauthorized third party [17]. The SSL protocol works using an SSL certificate inside the server, so when a customer requires access to it, it is done by means of a public key of the server making a safe connection by means of symmetrical cryptography. This validation is made by the browser web again without needing software installed. As long as, the quantum computing was not entirely developed this method of security joined with HTTP, meant, generating secure HTTPS versions “Fig. 6” that would keep being secure and used a lot in based cloud-based services.

Nonetheless, big companies that offer cloud services have already known the impact that the implementation of quantum computing can have, as too many cryptographic protocols will stop being secure. Because of this, companies like IBM, that have developed quantum computers with some qubits, had the idea of letting people access the quantum computer though the cloud [18]. As we mentioned before, it is too complex for homes or small businesses to have a quantum computer because extreme cooling is needed and due to any interference quantum properties might be lost. Thus, IBM opened the “IBM Quantum Computing” platform in which the community can register, program and generate quantum algorithms that are directly executed in a 16 qubits quantum computer. This is with the aim that programmers, scientists, including physicists, make their investigations tests using quantum computing and validate the efficiency and the answer speed. The mode to program on this platform is by the terms of quantum rational ports, as the logic port of Hadamard, phase displacement, the SAWP gate, CNOT, and Controlled-U, amongst others “Fig. 7”. These logic gates are moved as pieces inside a virtual circuit over each qubit line as is shown in the following chart.

Besides, there is a platform that lets one create quantum algorithms and execute them in its computer. IBM is developing a service in the cloud of quantum cryptography to let several companies requiring to increase their internet

security to do so. They could do so by connecting to these services from the year 2020. Of course, access to these services is awfully expensive but it guarantees that the security of the information will be safer, despite quantum computation being implemented.

Also, Microsoft with its cloud platform, “Azure”, is offering quantum services that provide free a group of tools for quantum programming called Q# (Q-Sharp) of open source and available in GitHub. It has a package of quantum solutions precompiled ready to be used in the projects inside Azure, also including simulators to test the algorithms before executing them in the quantum real machine. Equally, it is expected the giant Google will open a platform to the public to access its quantum servers as they have been recognized to reach quantum supremacy by executing an algorithm of random numbers in just 200 seconds. A normal computer would take nearly 10,000 years [19]. Amazon, another of the big technology companies, also opened a platform in the cloud called “Braket”, which offers quantum computation as a cloud service [20]. This allows the creation of a community in which they only develop quantum algorithms to be tested afterwards by the Amazon simulator, AWS. They offer an environment of work which is, *Jupyter*, entirely managed. All these tools of Amazon allow the investigation and identification of applications of quantum computation that can be feasible for the companies’ customers.

As has been observed, the big technology companies are opening their doors using the cloud so that those interested in the subject can collaborate in the development of quantum computation. The cloud acts as a bridge between the programmers, scientists, physicists, amongst others and the quantum computer [21]. Then, a company that provides services on the internet, like on-line music, can start up the investigation of how to improve its security using these platforms of quantum processing, generating new algorithms or improving the cryptography they are using. The problem in using these platforms from third parties is that we continue depending on their quantum computers. By their hardware they can take advantage of this technology by implementing methods of collection in their services. However, it is not a problem for those that really wish to increase their security in the cloud, like in the case of the banking agencies and virtual shops who could develop software that connects to these quantum services to prevent fraud in electronic transactions.

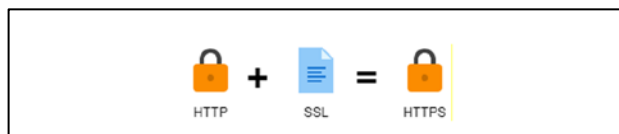


Fig. 6. Https Protocol.

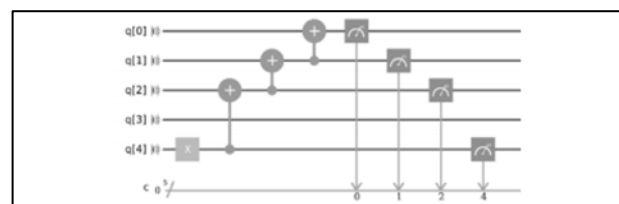


Fig. 7. Quantum Logic Gates.

That said, one of the important properties of quantum physics is quantum teleportation which takes advantage of quantum entanglement to send information from place to place. In 1997 the experiment to transport information was conducted on separate quantum particles to less than a meter. They did it and realized that this could be duplicated, even though the quantum particles were more widely separated. In 2012 a team of physicists from the Austrian Academy of the Sciences of the University of Vienna in Austria carried out quantum teleportation to 143 kilometers successfully [22]. From these developments they have made longer distance teleportation from space and underwater to 1,200 kilometers, quantum chips entangled sending information instantly without existing physical or electronic connection between both circuits. It is said, that if in a future a network or quantum cloud existed these servers could send information without needing a physical connection between them. Theoretically speaking it would be enough only to take advantage of quantum entanglement; this would suppose an exceedingly high security in the quantum communications as an interceptor or hacker does not have a channel that intercepts. The problem with this is that it only happens between quantum servers and not in compound networks. However, if the technology continues advancing so quickly in maybe 50 years or more there could be more quantum computers forming a new cloud between them [23].

III. DISCUSSIONS

Information security is an essential subject in telecommunications. If security did not exist its use would be extremely limited as we could not create online stores, there would be systems of information with session start, and we could not make bank transactions, amongst other things. Therefore, security always will be involved in a continuous study where security methods are increasingly difficult to break and the information is kept safe. For this reason, different algorithms have been created to allow a secure information transmission mode; algorithms like the AES, GIVE and MD5 have taken enough strength and are practically every electronic device like cellphones, computers, credit cards, TVs, cameras of surveillance cameras, etc. [24]. The mathematics that these algorithms use is modular mathematics, factorization of prime numbers, combinations, permutations etc., thus allowing a greater complexity for a third party or attacker trying to break the method of encryption. This has been possible due to the computing power that has been evolving through time. Cryptology has taken advantage of the computational power increasing the quantity of bits to be used to encipher a message. Even so, the same computational systems are unable to break systems of security in a short time but can take many years processing and trying to discover the keys that allow decoding of the information.

Then, why is a piece of information not completely secure is because its channel of communication can be easily intercepted or because it does not use big enough keys to encrypt the information. Therefore, it is necessary to consider that its channel should be the most private possible. Attackers sometimes use social engineering so that the people who are sending or receiving confidential information are unaware and show the key that allows decoding of this information., It may be by just showing fake advertising in web pages, fake buttons

or application of information. Also, the intruder through these techniques can insert malware which allows control of the computer or simply collection of information. Thus, when it requires greater security requires greater computational using more power space and more hardware. This does not mean that it is a negative benefit to improve computational power with the aim of improving information security. On the contrary, it is a positive by improving the methods of encryption making us feel safer sharing information or making bank transactions. Yet this is not entirely sufficient. The current computational power is based in binary systems, with all the calculations, processing and digital life based in the binary system which when it creates a supercomputer is able to process much faster mathematical complex calculations than a conventional computer. We can say that a system of encryption is in danger of being decoded in a shorter time. Current supercomputers like the Summit of IBM instrumented with artificial intelligence have a power of 200 petaflops [25], this means, that they can do 200 floating comma operations per second and could be able to decode some algorithms of cryptography.

Now, quantum physics has taught us that it can be used to create quantum computers where we take advantage of the main characteristics of these physical states, which means if a conventional computer based in binary systems (1 or 0), the quantum computer can take the 1 and 0 at the same time, being a lot more powerful, faster than a supercomputer and without occupying a lot of space. If a supercomputer based in the binary system can break some cryptographic systems, the quantum computer also will do it in less time. An example of this is the publication by Google where its quantum computer was able to execute an algorithm that generates random numbers in 200 seconds when a conventional supercomputer can take 10.000 years. Even up to now some studies as well as prototypes of quantum computers have been made [26]. The scope of these computers is quite wide, having important uses in medicine, astronomy, chemistry, physics and of course in computing especially in the field of cryptography, where it has designed quantum algorithms that allow a research of information in non-orderly data (called the algorithm of Grover). This means data does not even require to be organized in intelligible form for informational research. Another algorithm called the Algorithm of Shor performs factorization of numbers using quantum physics. This is important in security since many of the current cryptographic algorithms are based in breaking down a number to prime factors. It means with the implementation of quantum computation all the current systems of security would be in danger, as a quantum computer with sufficient capacity of Qubits, [the basic unit of the quantum computer as the bit is for the conventional computer], can easily break these algorithms.

Thus, they have produced diverse studies in quantum cryptography. In addition, protocols that allow one to perform a safe communication like the protocol EPR taking advantage of entangled quantum between pairs of photons. This means that when two photons are interwoven, and a photon changes the state or makes some perturbation immediately the other photon takes on the same behavior. This property is now particularly useful to send information and in a safe mode, since with a small perturbation that comes from no controlled

external entities one can modify the original information warning the receptor that the message has been intercepted and is not intelligible at the moment of decoding. If we take advantage of this property of quantum physics in quantum computers, it will improve security significantly since the intruder could not try to read the message because it would be modified without intention and immediately raise an alarm indicating an intervention in the communication. However, the same was thought when they created the first systems of cryptography like the enciphered Caesar. There it was believed for decoding it was necessary to use black magic or ask the help of some witchcraft. Also, with the enigma machine that was used by the Nazis it was thought that it was an impossible system to decode until Alan Turing managed to do it due to his knowledge of mathematics and logic; then also it can succeed in finding the way to intercept the messages in the quantum computers not immediately but when these quantum theories of the physics and quantum mechanics are near to be in the use a lot of people who can study and understand them. Therefore, the competition between them which is apt to create the best unbreakable encrypted algorithm for the coder and the attacker who wants to decode it in the shortest time will continue existing even with the arrival of quantum computation.

For quantum computation use to be near all of us, as well as the current computation, a lot is missing for implementation since the existing quantum computers suffer from different problems. Firstly, the quantum computer requires quite a lot of refrigeration that can be costly as the particles being in a quantum state generate quite a lot of heat. Secondly, the current quantum computers still do not have the necessary power to completely replace the current ones. The first quantum computers would be used especially by scientists in matters of biology, astronomy, chemistry, amongst other. There does exist the possibility of accessing a quantum computer via the cloud for academic use, thanks to platforms launched by IBM, Amazon, and Microsoft. They can design quantum algorithms with new programming languages and test their functionality. It is here where the community that devotes to the cryptology can investigate and develop new methods of computer security that in the future can be applied in networks of quantum computers which require to communicate in a secure mode. Meanwhile the cloud can generate something as well as "quantum services" that are to the order of those that require quantum computational power for some internal process of some company. For example, a programmer of a security company can develop an algorithm that strengthens the system of security that is internally in the company. A banking entity can improve its security of transactions connecting its system of security to one of these quantum services. A software app for text mailing like WhatsApp or Telegram can make use of these quantum services to improve its security in sending messages. Software of streaming like Netflix or HBO can avoid theft or copies from original series. Software of VPN can improve its efficiency taking advantage of these quantum services to improve the privacy and security in those companies that use VPN for safe connections in the web. So, with this successively different software, banking and governmental companies can take advantage of this new technology.

Taking advantage of this new technology depends on the freedom to access the cloud of a quantum computer. Since its use can in future carry additional cost, for example, as well as Microsoft Azure, different services that are by payment, will also carry quantum services, and is not too surprising since their price is necessary to keep a cutting-edge technology progressing obtaining a favorable result for the customers. It is true that, if very few know their potential, demand for consumption of these services will be low. Therefore, it is important that academies of computer security even universities or schools begin to explore the world of quantum computation. The programming of quantum computers that already exists varies like the Q# of Microsoft, Quipper, the Quantum Composer of IBM, amongst others; as professionals exist in the subject they can develop the new cryptographic algorithms that improve technological security in the world. Since a company can apply, into its systems of small security, a process to connect into a quantum service, they will see the success that it has by observing that the system is more reliable and faster. It will be popular; it will attract a lot of more customers and its income will be much higher; from this moment other companies will be attracted by this technology and surely incorporate it in their services. It is here where a new need opens to purchase the knowledge of programmers, scientists and physicists in the quantum subject. Therefore, it is important to begin to learn to use this technology that is for the future.

Another field that will importantly be affected by quantum cryptography are the cryptocurrencies [27]. Although it is known that cryptocurrencies save a unique hash that is not able to be copied and their system of transactions have security, therefore it has a value. When we dominate the quantum cryptography the majority of cryptocurrencies will no longer be safe and therefore its use would go down use since it is a totally digital coin. However crypto coins could also see benefit if they inject quantum cryptographic algorithms into their systems of security algorithms quantum, as is done already by Bitcoin, Ethereum. Or possibly develop a new cryptocurrencies based in quantum computation, that in my opinion would make cryptocurrencies safer and their use would increase. This subject of security where money is at stake is important for banking agencies which would have to be concerned not only because their technological systems would become vulnerable but also because if the quantum cryptocurrencies arrived first before of the agencies increased their security, the use of banks would lessen.

IV. CONCLUSIONS

This new technology like any other, has to be controlled. They cannot have all their potential immediately belonging to all community as it could be used for hacking different systems. The first step is to allow important sectors, like banking, and government programs to use this technology with all its potential, as they improve its system against attacks and afterwards put at the disposal of the public a small part of the power that quantum computation would have as we know that the systems of current security can be very vulnerable.

ACKNOWLEDGMENT

The authors would like to express their gratitude to the faculty of technology of the Universidad Distrital Francisco Jose de Caldas. From Bogotá, Colombia.

REFERENCES

- [1] Domínguez Espinoza, Edgar Uriel. Pacheco Gómez Leonardo. Classic cryptography algorithms. apr 10th 2007. National Autonomous University of Mexico. [Quoted apr 23rd 2019]. Available at: <https://docplayer.es/21024045-Universidad-nacional-autonoma-de-mexico-facultad-de-ingenieria-criptografia-algoritmos-de-criptografia-clasica.html>.
- [2] Héctor Corrales Sánchez, Carlos Cilleruelo Rodríguez, Alejandro Cuevas Notario. Cryptography and Encryption Methods. Madrid. apr 24th 2014. University of Alcalá. [Quoted apr 25th 2019]. Available at: <http://www3.uah.es/libretics/concurso2014/files2014/Trabajos/Criptografia%20y%20Metodos%20de%20Cifrado.pdf>.
- [3] B. Bhat, A. W. Ali and A. Gupta, "DES and AES performance evaluation," International Conference on Computing, Communication & Automation", Noida, 2015, pp. 887-890. doi: 10.1109/CCAA.2015.7148500.
- [4] Tao Xie, Fanbao Liu, Dengguo Feng. Fast Collision Attack on MD5. 2013. [Quoted apr 23rd 2019]. Available at: <https://pdfs.semanticscholar.org/a9d4/833698895915d34f2ac5509f1bf0887b4c5b.pdf>.
- [5] D.J. Griffiths. Introduction to Quantum Mechanics. Prentice Hall - New Jersey. 1995. [Quoted apr 15th 2019]. Available at: http://gr.xjtu.edu.cn/c/document_library/get_file?p_l_id=21699&folderId=2383652&name=DLFE-82647.pdf.
- [6] Ahmed Banafa. Understand quantum cryptography. nov 19th 2015. BBVAOPENMIND. [Quoted may 9th 2019]. Available at: <https://www.bbvaopenmind.com/tecnologia/mundo-digital/entender-la-criptografia-cuantica/>.
- [7] Quantum cryptography - Cryptography concepts. jun 24th 2005. Scientific texts. [Quoted may 9th 2019]. Available at: <https://www.textoscientificos.com/criptografia/quantica>.
- [8] Nithin Nagaraj, Vivek Prabhakar Vaidya, Prabhakar Govind Vaidya. Re-visiting the One-Time Pad. 2008. GE Global Research. [Quoted apr 28th 2019]. Available at: https://www.researchgate.net/publication/220284469_Re-visiting_the_One-Time_Pad.
- [9] Javier Blanco. Quantum computers could break RSA 2048-bit encryption in 8 hours?. dec 29th 2019. Agency6. [Quoted jan 30th 2020]. Available at: <https://agencia6.com/index.php/2019/12/29/ordenadores-cuanticos-podrian-romper-el-cifrado-rsa-de-2048-bits-en-8-horas/>.
- [10] Leander Kahney. Quantum Leap in Searching. jul 3rd 2011 [Quoted jun 4th 2019]. Available at: <https://web.archive.org/web/20110703044742/http://www.wired.com/science/discoveries/news/2000/05/36574>.
- [11] Z. Sakhi, R. Kabil, A. Tragha and M. Bennai, "Quantum cryptography based on Grover's algorithm," Second International Conference on the Innovative Computing Technology (INTECH 2012), Casablanca, 2012, pp. 33-37, doi: 10.1109/INTECH.2012.6457788.
- [12] David Moreno Centeno. Post-quantum cryptography: McEliece implementation and a new version. University of Valladolid. [Quoted feb 19th 2020]. Available at: <https://uvadoc.uva.es/bitstream/handle/10324/38361/TFG-B.1368.pdf?sequence=1&isAllowed=y>.
- [13] Cryptography in the financial world. sep 2nd 2014. Media-Tics Information and communication in the digital age. [Quoted oct 3rd 2019]. Available at: <https://www.media-tics.com/noticia/2585/blogs/la-criptografia-en-el-mundo-financiero.html>.
- [14] A. Einstein, B. Podolsky and N. Rosen. "Can Quantum-Mechanical Description of Physical Reality be Considered Complete?". Physical Review. Vol. 47 N° 10, pp.777-780. May 15th of 1935. DOI: 10.1103/PhysRev.47.777.
- [15] Javier Paniagua. Cryptography in banking. jul 7th 2017. Research IT Now. [Quoted may 7th 2020]. Available at: <https://revistaitnow.com/criptografia-en-la-banca/>.
- [16] Steve Jurvetson. A quantum computer will break RSA 2048-bit encryption in eight hours. jun 9th 2019. Source Sites. [Quoted may 21th 2020]. Available at: <https://sitiosfuente.info/ciencias/13147-ordenador-cuatico-cifrado.html>.
- [17] Mónica Tilves. IBM Cloud to deliver quantum security cryptography services starting in 2020. aug 26th 2019. Silicon. [Quoted jun 4th 2020]. Available at: <https://www.silicon.es/ibm-cloud-entregara-servicios-de-criptografia-de-seguridad-cuantica-a-partir-de-2020-2402341>.
- [18] IBM to offer quantum-safe encryption services over public cloud in 2020. aug 26th 2019. Digital Security. [Quoted nov 14th 2019]. Available at: <https://www.itdigitalsecurity.es/cloud/2019/08/ibm-ofrecera-servicios-de-cifrado-quantumsafe-sobre-nube-publica-en-2020>.
- [19] Josep Corbella. Google proves quantum supremacy. oct 23th 2019. [Quoted nov 14th 2019]. Available at: <https://www.lavanguardia.com/ciencia/20191023/471156519790/ordenador-cuatico-google-supremacia-computacion-cuantica.html>.
- [20] Celia Valdeolmillos. AWS re:Invent 2019: Amazon bets on quantum computing with Bracket. dec 3rd 2019. [Quoted feb 2nd 2020]. Available at: <https://www.muycomputerpro.com/2019/12/03/aws-reinvent-2019-amazon-computacion-cuantica-braket>.
- [21] Lara Olmo. Google sees new possibilities for its cloud business in quantum computing. jul 17th 2017. [Quoted nov 10th 2019]. Available at: <https://www.ticbeat.com/tecnologias/google-ve-en-la-computacion-cuantica-nuevas-posibilidades-para-su-negocio-cloud/>.
- [22] New record in quantum teleportation lays the foundation for global quantum communication. sep 6th 2012. [Quoted oct 10th 2019]. Available at: https://tendencias21.levante-emv.com/nuevo-record-en-teleportacion-cuantica-sienta-las-bases-para-una-comunicacion-cuantica-global_a13021.html.
- [23] Llewellyn. Powerful boost to the quantum Internet. jan 3rd 2020. [Quoted mar 13rd 2020]. Available at: https://tendencias21.levante-emv.com/potente-impulso-al-internet-cuatico_a45631.html.
- [24] A. K. Mandal, C. Parakash and A. Tiwari, "Performance evaluation of cryptographic algorithms: DES and AES," 2012 IEEE Students' Conference on Electrical, Electronics and Computer Science, Bhopal, 2012, pp. 1-5.
- [25] The world's most powerful supercomputer identified chemicals that could stop the spread of the coronavirus. mar 21st 2020. Infobae. [Quoted apr 4th 2020]. Available at: <https://www.infobae.com/america/mundo/2020/03/21/la-supercomputadora-mas-potente-del-mundo-identifico-los-quimicos-que-podrian-detener-la-propagacion-del-coronavirus/>.
- [26] Google publishes how it has achieved quantum supremacy. oct 23rd 2019. [Quoted apr 4th 2020]. Available at: <https://www.publico.es/ciencias/ordenador-cuatico-google-publica-logrado-supremacia-cuantica.html>.
- [27] Cassio Gusson. Bitcoin and Ethereum developers prepare for quantum computers. oct 12nd 2019. [Quoted apr 4th 2020]. Available at: <https://es.cointelegraph.com/news/bitcoin-and-ethereum-developers-prepare-cryptocurrencies-for-quantum-computing>.

Energy Storage and Electric Vehicles: Technology, Operation, Challenges, and Cost-Benefit Analysis

Surender Reddy Salkuti

Department of Railroad and Electrical Engineering, Woosong University, Daejeon, Republic of Korea

Abstract—With ever-increasing oil prices and concerns for the natural environment, there is a fast-growing interest in electric vehicles (EVs) and renewable energy resources (RERs), and they play an important role in a gradual transition. However, energy storage is the weak point of EVs that delays their progress. The world's EV industry is accelerating to faster adoption with appropriate incentives to the EV owners, policy support, and encouraging local manufacturing. The increasing demand for EV's has presented itself as an authentic alternative to internal combustion engines (ICE). The main feature of the RERs is their variability and intermittency. These drawbacks are overcome by integrating more than one renewable energy source including backup sources and storage systems. This paper presents various technologies, operations, challenges, and cost-benefit analysis of energy storage systems and EVs.

Keywords—Energy storage; electric vehicles; cost-benefit analysis; demand-side management; renewable energy; smart grid

I. INTRODUCTION

The demand for the electrical energy is increasing in the modern world; however the fossil fuel-based energy systems are polluting and depleting existing the available reserves. Environmental awareness is worldwide increasing. New paradigms are emerging, like electric vehicles (EVs), smart grids, electrical markets, and vehicle-to-grid (V2G). The novel grid techniques are demonstrated for the optimal integrated operation of RERs and EVs to increase the penetration of renewable energy. The need for conservation of fast-depleting natural resources and concerns about environmental protection are demanding for sustainable green energy technologies [1, 2]. Nowadays, the focus on alternative renewable energy is increased and the predominant RERs are wind power, solar PV power, and hydro power. The amount of power availability from these RERs follows some daily and seasonal patterns, but the power demand by the consumers follows different characteristics. Solar PV generation is omnipresent, low operational cost, less maintenance, it can be easily accommodated in roof-tops and eco-friendly energy conversion processes. The integration of RERs makes the system more complex concerning the power-sharing, control, and analysis [3].

The international energy roadmap study ranks solar PV, biomass, windmills, and tidal power as future sources of renewable energy to sustain the world's economy. Progress in the field of sustainable energy scenario over the previous span of time has been exceptional. The two main sustainable energy resources are the sun and the wind [4]. The growth in electricity generation from renewable was substantially increased. Toward the finish of 2016, the world's aggregate

wind power generation remained at 486.8 GW. 2016 has seen near to 75 GW of extra introduced PV capacity around the world, a stunning 50% over 2015 and raising the aggregate newly introduced capacity to around 300 GW [5]. As in 2016, China, USA and Japan displayed the biggest markets representing seventy-five percent of the extra introduced capacity in these three nations alone [6, 7]. In the interim, 66% of the worldwide PV limit is being introduced in the Asia-Pacific locale with China in front of all others over 34 GW of installed limit in 2016. 24 nations have now achieved total installed limits over 1 GW, 16 nations introduced no less than 500 MW amid 2016 and in no less than 27 nations, PV contributes with 1% or more to the yearly power supply. In 2017, PV will add to around 2% of the world's power generation [8, 9]. The IEA ventures by 2050, around (15-18) % of worldwide power is going to be created from the wind with sun-powered PV contributing as high as 16% [10, 11] regardless of its different favorable circumstances.

With the integration of RERs such as wind and solar, significant uncertainty into the power system is developed. It is a great challenge for system operators to maintain reliable operation and efficient electricity markets with simultaneous maximum utilization of renewable energy. As the electric market structures change to improve the management of renewable sources, advances in the design and pricing aspects of energy and ancillary services markets are needed. The main objectives of an energy management system are to ensure the maximum utilization of RERs, continuous power supply to the load, reduce the cost of energy production and increase the stability of the system [12]. To achieve these objectives efficiently and fast control techniques are required which are capable of processing information intelligently and taking critical decisions dynamically within the operational constraints.

EVs are propelled by electric motors and use the electrical energy stored in the batteries. EVs are required to reduce the dependence on fossil fuel and to reduce pollution as transportation accounts for one-third of all energy usage. By using the EVs 100%, the CO_2 emission can be reduced by half. EVs save energy, less pollution, and noise, cheaper to run and maintain. However, they also include some challenges such as selecting the battery size and its capacity, locations of charging stations, faster charging, speed and mileage of the vehicle, and its efficiency. Various components of EVs include electronic controllers, energy storage systems, power electronic converters, and electric motors [13]. Various EV technologies include battery EVs (BEVs), hybrid EVs (HEVs), plug-in HEVs (PHEVs), and fuel cell EVs (FCEVs).

The author in [14] reviews the recent trends and directions of optimal control strategies of ESSs. The author in [15] presents the potential profit by using the second-life batteries from the EVs. A comparison between various integration approaches of EVs subjected to the availability of PV systems and superconducting magnetic ESSs has been presented in [16]. The author in [17] describes the energy management benchmarks and sizing guides of battery-supercapacitor hybrid ESSs in EV applications. The author in [18] presents different EV distributed renewable energy coordination approaches by considering the costs and the associated infrastructure. Analysis of energy storage tanks and the types of accumulators used for EVs and the patterns of the Li-ion battery is presented in [19]. The author in [20] presents the estimation of supercapacitor storage influence on Li-ion battery cycle life and EV performance. An overview of various technologies of ESSs, their characteristics, constructions, classifications, evaluation, and electricity conversion processes with advantages and disadvantages for EV applications has been presented in [21].

The integration of renewable sources such as wind and solar introduce significant uncertainty into the power system. It is a great challenge for system operators to reliable operation and efficient electricity markets with simultaneous maximum utilization of renewable energy. As the electric market structures change to improve the management of renewable sources, advances in the design and pricing aspects of energy and ancillary services markets are needed. In this paper, the integration of storage devices including the existing storage technologies such as pumped hydro as well as utility-scale battery systems collocated with solar and wind farms are considered. Due to the high capital cost of the energy storage systems, a study is performed considering the trade-off between the economic costs and reliability for different levels of penetration of these systems. The main objective of this work is to determine suitable ways to combine some forms of energy storage systems (ESSs) such as flywheel with lithium-ion batteries to achieve load balancing in the smart grid.

The remaining work of this paper is prepared as follows: Section II presents the description of various energy storage systems. Different electric vehicles are described in Section III. Section IV presents the cost-benefit analysis. Conclusions are summarized in Section V.

II. ENERGY STORAGE SYSTEMS (ESS)

RERs are unpredictable and there is a gap between the availability and usage of these resources. There is always a difference between the production and consumption of such resources. As a result, it is of immense importance to build storage units that can preserve the energy and make it available for later use. In the present scenario, certain storage technologies include compressed air, supercapacitors, and advanced battery systems. Proper utilization of RERs will enable the protection and longevity of the environment by reducing pollution. They ensure continuity of energy supply and improve the reliability of the system by providing excellent energy management techniques. Energy storage systems can be in many forms and sizes. Energy can be stored as potential, kinetic, chemical, electromagnetic, thermal, etc.

[22, 23]. Some energy storage forms are better suited for small-scale systems as well as for large-scale storage systems. Some of the energy storage systems are chemical batteries, fuel cells, ultra-capacitors or supercapacitors, superconducting magnetic energy storage, and flywheels, etc. The potential applications of energy storage systems include utility, commercial and industrial, off-grid, and microgrid systems. Renewables with energy storage can act as the baseload power source of a microgrid and reduce the use of fossil-fuel-based generators [24]. Energy storage is the conversion of unused energy at any given time into a form that can be stored for use at a later time. The issue of energy storage arises with the need to match the demand and supply of energy to individuals. The advent of electricity brought about more concern for the need for energy storage due to its prior nature of being used up when generated or converted to another form of energy [25]. However, new trends in energy show ways these generated energies could be stored and harnessed.

A. Battery Energy Storage

A battery is an electrochemical device that stores electrical charges through chemical reactions. There are two types of batteries, primary (disposable) and secondary (rechargeable). A primary battery converts the chemical reactions into electricity only once. Batteries used for large-scale ESSs are secondary/rechargeable batteries. The charging and discharging of a battery is a reduction-oxidation process. During the discharge, electrons are transferred from the battery to the load through the process of oxidation [26]. When charging, electrons are transferred to the battery when a voltage is applied to its terminals. This is referred to as a reduction process.

Various battery energy storage technologies used for EVs include Lithium-ion, Lead-acid, Nickel-metal hydride, and Sodium nickel chloride. The first three batteries operate at room temperature whereas the last one operates at 300°C. A lithium-ion battery is a leader among battery storage technology for EVs. Sodium nickel chloride is a low maintenance battery with limited use as it is going to operate at 300°C. Battery-operated EVs have low initial infrastructure cost, low noise, high efficiency, low operational cost, and zero emissions [27]. Major challenges involved with battery technology are charge time, cost, shelf life, specific power, cycle life, specific energy, safety, ease of manufacture, and recyclability. Energy sources provide electrical energy onboard the EV. The types of energy storage technologies that have been proven to be viable and improvement have been going in are batteries (electrochemical cells), fuel cells, ultra-capacitors, and flywheel storage. In the foreseeable future, batteries are still the major source of energy for EVs. Newer types of batteries like metal-air, flow batteries, sodium-ion batteries, Sulphur based batteries, etc. [28]. The end goal is to achieve the same range as when gasoline is used in EVs. The battery energy storage system can be modeled as [29],

$$S_{t+1} = S_t(1 - \varepsilon) + [P_{DC}(t) + \eta_R P_{AC}(t) - P_D(t)]\eta_{CB} \quad (1)$$

$$S_{t+1} = S_t(1 - \varepsilon) - \left(\frac{\eta_I P_{AC}(t) - P_D(t)}{\eta_{DB}} \right) \quad (2)$$

Where S_{t+1} is state of charge of battery at time (t+1), ε is hourly self-discharge rate, $P_{AC}(t)$ is AC power generation at

time t , $P_{DC}(t)$ is DC power generation at time t , $P_D(t)$ is load demand at time t , η_R is the efficiency of rectifier system [30], η_{CB} is the charging efficiency of the battery, η_{DB} is the discharging efficiency of the battery, and η_I is the efficiency of the inverter system.

B. Lithium-ion (Li-ion) Batteries

Li-ion batteries are quite inexpensive, high energy, power densities, and highly efficient. This battery is composed of a negative electrode (carbon), a positive electrode (metal oxide), and an electrolyte (Lithium salt). These batteries do not pose a huge environmental impact. However, they tend to explode when exposed to high temperatures or short-circuited [31]. Flywheels are one of the oldest means of storing energy. It is a mechanical storage device that is used in storing rotational energy. They have a moment of inertia and resist the changes in rotational speed. It absorbs energy and acts as a reservoir. Torque is applied to a flywheel to transfer energy. Unlike flywheels, batteries are a slow chemical process that is subject to the recharge or discharge process [32]. Like all electrical devices, batteries can only recharge slowly regardless of the available input energy. Most flywheels are built to where they rarely wear out. The cost of batteries compared to flywheel storage is much cheaper.

C. Supercapacitors or Ultra-capacitors

These are the electromechanical capacitors that have usually high energy density when compared to common capacitors. For the same size as in conventional capacitors, the supercapacitors will have a capacitance of several farads, an improvement of about two or three orders of magnitude in capacitance, but usually at a lower working voltage, and hence they have very high energy densities [33]. It is having advantages over other technologies like a very high rate of charge and discharge, high output power, high efficiency, etc. Practical implementations of supercapacitors involve connecting various cells in parallel strings to maximize the storage capacity.

As the development of technology has merged towards EVs, the supercapacitors with high energy density are used for fast charging, and temperature stability. These supercapacitors are also used in flash photography devices, media players, automated meter reading. Supercapacitors have high efficiency (up to 95%), however, future research can be performed in terms of dielectric materials and other components by reducing the energy per unit weight. The charging rate of capacitors is limited by the current heating of an electrode. The layers in an electrical double layer are conductive when it is seen by itself, however, when they keep

contacting, then no current flow occurs. When supercapacitors are compared to the batteries, their energy density is about one-tenth of the battery. However, when it comes to power density, the supercapacitors have a higher rate which is about 10 to 100 times better than the battery [34]. The electrical double layer occurs when the two plates with the same condition are being separated and it leads to the separation of charges even though their separation possibility is very rare. The double-layer only can accelerate the low voltage but if it needs the higher voltage to be used, and then make the capacitors connected in series [35]. Various types of batteries for EVs and their specifications are presented in Table I.

D. Demand Side Management (DSM)

The conventional power grid is being replaced by a smarter grid through the implementation of several innovative changes such as demand response (DR) programs, distributed energy resources (DERs), and DSM. DSM is required because saving 1 unit of electricity at the consumer end avoids nearly 2.5 times of energy capacity addition. DSM leads to use less energy during peak hours, i.e., peak clipping and shifting the energy to use it in off-peak hours, i.e., valley filling. Therefore, DSM can help in meeting the demand at a lower cost. The DSM strategies include peak clipping, valley filling, load shifting, strategic conservation, strategic load growth, and flexible load shape [36]. The benefits of DSM reduce the peak load demand, reduces the cost of operation, minimizes the power import from the utility grid, and maximizes the use of RERs. Vehicle-to-grid (V2G) technology enables the energy to be pushed back to the power grid from the battery of an EV to help supply energy at the time of peak load demand. These EVs can provide a strong option with no extra cost. DSM acts as an energy-efficient measure that modifies/reduces the end user's energy demand, and it leads to cost reduction, environmental and social impact, reliability and network issues, and improved markets.

DSM provides energy efficiency options for the effective demand of customers and enabling them to cut their expenses by cost optimization. DSM also helps in the reduction of emissions and hence enables a more sustainable power system. It lowers the cost of transmission and also contributes to the reliability of generation. It obligates energy providers to maintain proper power levels and constant frequency levels. Moreover, with the advent of PV generation, DSM needs to be optimized further to enable it to handle fluctuation in power supply and system frequency, which may affect the operation of consumer appliances. In addition to that, optimal DSM technologies should also be developed for distributors and storage equipment.

TABLE I. VARIOUS TYPES OF BATTERIES FOR EVs AND THEIR SPECIFICATIONS

Name of the battery	Specific energy (Wh/kg)	Specific Power (W/kg)	Number of life cycles	Cost (\$/kWh)
Valve regulated lead acid	30-45	200-300	400-600	150
Nickel Cadmium	40-60	150-350	600-1200	300
Nickel Metal Hydride	60-120	150-400	600-1200	200-350
Zinc-air	230	105	N/A	90-120
Sodium-Sulphur	100	200	800	250-450
Lithium-ion	90-160	250-450	1200-2000	600-1000

III. ELECTRIC VEHICLES (EVs)

An EV is a mode of transport that is partly or completely propelled using electricity as its energy source. EVs were invented 178 years ago. The first EV was a battery-powered tricycle built by Thomas Davenport in 1834. EVs eventually lost the competition for dominance to combustion engines. Interests in EVs rekindled in the 1970s due to the outbreak of the energy crisis and oil shortage. The developmental pace of EVs accelerated when there was growing concern over air quality and the greenhouse effect in the 1980s. To achieve the desired level of performance, the ESSs (i.e., batteries) inside the vehicle need to have a very high capacity and efficiency to be able to give the vehicle enough power to drive. The development of more improved batteries currently plays a huge role in the design of EVs.

Apart from technological advancement, access to charging infrastructure plays a vital role in scaling up EV's. EV operation from the grid perspective includes load regulation with voltage and current constraints, maximize operational efficiency, and for EV aggregators perspective it manages the ancillary services, charging fairness, and minimizing the battery degradation. From the EV cost aspects, the objective of the problem includes the minimization of cost of power operations and EV charging costs and the maximization of grid operator revenue and aggregator profit. Various research problems that need to be addressed related to EVs are the demand response, selection of EV charging station, distribution system operation, frequency regulation by EV benefits maximization of profit of parking lot owner and vehicular energy network, etc. [37]. From the grid point of view, the challenges in EVs include infrastructure, power quality, and intermittency as EV is an intermittent load and includes large penetration of RERs and grid operators to maintain grid stability. For the analysis of EVs, the data required includes the energy supply data, demand data, road network data, power network data, forecast data, load curve with and without EVs, and the time of day prices.

A. Classification of EVs

EVs are mainly classified based on their energy sources and the propulsion devices as battery electric vehicles (BEVs), hybrid electric vehicles (HEVs), fuel cell electric vehicles (FCEVs), and plug-in hybrid EVs (PHEVs). A hybrid EV has two or more power sources and there are a large number of possible variations. These hybrid EVs combine an internal combustion engine (ICE) with a better and electric motor and generator. These hybrid EVs are maybe in series hybrid or parallel hybrid.

B. Battery Electric Vehicles (BEVs)

The BEVs are generally termed EVs, rely on using batteries as their sole or major source of the energy storage device to store electricity. They have minimum overall emissions and zero tailpipe emissions. At the present status of battery technology, the energy storage capacity of BEVs is far less than that of internal combustion engine vehicles (ICEVs). Therefore they have a problem of range anxiety and also, BEV's are more expensive than general ICEVs. There is also the factor of charging time, for the batteries which depends on

the previous state of charge of the batteries and their size [38]. But due to the recent introduction of fast charging methods and/or battery swapping, the charging time problem can be solved.

C. Hybrid Electric Vehicles (HEVs)

HEVs implement an ICE combined with an electric motor, where the participation of the electric motor in the net power produced decides the HEV grade (micro, mild and full). HEVs have improved fuel economy, lesser environmental impact than their ICEV counterparts, and longer range too. But these vehicles become costlier due to their extra inventory.

D. Fuel Cell Electric Vehicles (FCEVs)

These vehicles also offer zero-tailpipe emissions and minimum overall emissions. Besides, the driving range in these vehicles is comparable to that of an ICEV. Its major challenges are the lacking technology to produce safe fuel cells, and they have a very high initial cost. There is also a lack of hydrogen refueling infrastructure, which would take a huge investment cost to set up. The commercialization and mass production of FCEVs in the future will depend on whether there will be a technological breakthrough in fuel cell technology.

E. Plug-in Hybrid EVs (PHEVs)

In PHEVs, the storage batteries are charged from regenerative braking energy, external charging, and by an engine connected through the generator, however, the amount of energy received from regenerative braking is very less. Therefore, for the driving mode of operation of EV, an engine connected through a generator is used [39]. They are designed to extend the range of an EV without compromising the performance, and therefore the cost is high.

F. Motor Drive EV Technology

Motor drives convert the on-board electrical energy to the desired mechanical motion. The motor drive technology for EVs is being rapidly improved in recent years, due to the developments in the design, analysis, and control of motor drives. Electric machines are the key elements of motor drive technology. The performance requirements of electric machines for EVs are much more demanding than those for industrial applications. The requirements of the electric motors include high torque and power density, wide constant power operating capability, high reliability and robustness, high efficiency over wide torque and speed ranges, high torque capability to compensate acceleration variations (start and climbing), cost efficiency, and low acoustic noise and losses.

EV machines are classified into commutator and commutator-less machines. The latter doesn't have a commutator or brushes. Motors like DC motors, permanent magnet DC (PMDC) motor, cage rotor induction motors, permanent magnet brushless (AC and DC), and switched reluctance motors (SRMs) have been widely applied to EVs. There is also a trend of developing new types of doubly salient configurations for EV motors like the doubly salient permanent magnet (DSPM) [36]. The advantages of using brushless DC (BLDC) motors include low maintenance, 90+% efficiency, high operating speeds, no brush sparking, and

compact size, quick response, less rotor inertia, predictable speed regulation, quieter operation, and regenerative braking with good efficiency. As one can notice, the benefits of using BLDC motors are higher especially from its brushed counterparts and induction motors, making it more suitable for EV applications.

G. Challenges and Issues of EVs

Various challenges and issues associated with EVs are: EVs should be available at a competitive price when compared with internal combustion engines (ICE), the efficiency of EV's should be more at high speed and high load, quality of vehicles, safety and security features during usage. Other challenges include the import of battery manufacturing consists of lithium, nickel, cobalt, aluminum, and graphite, dependency and price volatile may impact EVs' total cost, the impact of EV charging stations on the grid, which create huge demand for electric power and may cause additional burden and unscheduled load demands, providing reliable charging infrastructure and determining the cost of charging, availability of land in densely populated urban localities for charging stations, and dynamic, predictable and encouraging government policies are needed for creating a necessary ecosystem of EVs.

IV. COST-BENEFIT ANALYSIS (CBA)

The cost-benefit strategy/analysis for quantifying the impact of RERs penetration with the consideration of PV/optimum tracking problem of renewable energy mix and control signals to the network performance has been presented in this paper. Ranking of controls would be done to compute the impacts in maintaining secure and reliable power management and the development of integrated schemes for real-time power management with RERs [32]. The unit commitment (UC) problem is an electrical generation unit scheduling problem. The aim is to determine a schedule over a time horizon that satisfies forecasted demand and technical requirements. Historically integer programming and other optimization approaches have been used to solve the UC problem. Additional challenges for solving the UC include the integration and optimization of very high, variable RERs and flexible consumer demand. Stochastic and robust optimization techniques seek to address the significant variability in the problem. To develop the criteria for affordability, reliability constraints, and economic benefits; the computation of price-based uncertainties must be developed by using advanced optimization schemes with foresight and flexibility. The CBA of penetration of RERs into RER integrated system in terms of impact on network performance must be evaluated and computed.

World's energy roadmap study ranks PV, biomass, windmills, and tidal power as future sources of renewable energy to sustain the world's economy. Traditional studies on RERs integration have focused on power quality improvement using one or two RERs, and on their impacts on grid performance. The components that made up the RERs integrated systems and RERs would be energy storage, load control, and advanced power electronics, which interface between the RERs and the grid provider. The electricity demand is rarely constrained over time. This task involves the

review of conventional and innovative storage methodologies for incorporation and optimal use of the RERs energy outputs. A study on modeling and performance of both conventional storage methodologies (i.e. batteries), innovative storage (i.e. flywheels) methodologies, associated power electronics for conversion and control, and an introduction of storage technologies to an integrated environment for simulation and experimentation will be useful for understanding this analysis [38]. For this, the capability of optimized and integrated renewable energy management systems using real-time measurements and local control is required. Depending on the power system an engineer desires to build, the CBA can help select a storage option. For example, if discharge duration was the highest on an engineers' priority list, then choosing the storage technique that allows getting the maximum benefit (discharge duration) for the given budget would be the selection.

V. CONCLUSIONS

This paper presents the need for electric vehicles (EVs) and various critical aspects that are instrumental in making EVs a sustainable, reliable, and affordable option for future mobility. Due to the intermittent nature of renewable energy resources (RERs), it is of utmost importance to store the excess energy when available and can be used for periods of excess load or inefficient supply. Due to the growing awareness of the harmful impact of conventional fossil fuels and advancements in renewable energy technologies, energy storage and EVs have grown in popularity. In this work, conventional and innovative storage techniques such as batteries, supercapacitors, and EVs are reviewed along with their merits, limitations, and performance. This paper also presents the cost-benefit analysis which is helpful to make a good engineering decision. Having strong numerical evidence based on cost is a great way to invest.

ACKNOWLEDGMENT

This research work was funded by "Woosong University's Academic Research Funding – 2021".

REFERENCES

- [1] S. Sharma, A.K. Panwar, M.M. Tripathi. Storage technologies for electric vehicles, *Journal of Traffic and Transportation Engineering*, vol. 7, no. 3, pp. 340-361, Jun. 2020.
- [2] X. Wu, X. Hu, X. Yin, S.J. Moura. Stochastic Optimal Energy Management of Smart Home With PEV Energy Storage, *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2065-2075, May 2018.
- [3] M. Stecca, L.R. Elizondo, T.B. Soeiro, P. Bauer, P. Palensky, A Comprehensive Review of the Integration of Battery Energy Storage Systems Into Distribution Networks, *IEEE Open Journal of the Industrial Electronics Society*, vol. 1, pp. 46-65, 2020.
- [4] C. Yang, M. Zha, W. Wang, K. Liu, C. Xiang. Efficient energy management strategy for hybrid electric vehicles/plug-in hybrid electric vehicles: review and recent advances under intelligent transportation system, *IET Intelligent Transport Systems*, vol. 14, no. 7, pp. 702-711, 7 2020.
- [5] F. Nadeem, S.M.S. Hussain, P.K. Tiwari, A.K. Goswami, T.S. Ustun. Comparative Review of Energy Storage Systems, Their Roles, and Impacts on Future Power Systems, *IEEE Access*, vol. 7, pp. 4555-4585, 2019.
- [6] X. Chen, X. Zhang. Secure Electricity Trading and Incentive Contract Model for Electric Vehicle Based on Energy Blockchain, *IEEE Access*, vol. 7, pp. 178763-178778, 2019.

- [7] Modeling, Control, and Integration of Energy Storage Systems in E-Transportation and Smart Grid, IEEE Transactions on Industrial Electronics, vol. 65, no. 8, pp. 6548-6551, Aug. 2018.
- [8] Opening up new markets for business, Global wind energy council, Technical Report, 2015. Available. [Online]. https://www.gwec.net/wp-content/uploads/vip/GWEC-Global-Wind-Report_2016.pdf.
- [9] PVPS annual report 2016, Technical Report. Available. [Online]. https://iea-pvps.org/wp-content/uploads/2020/01/IEA-PVPS_RA2016-web.pdf.
- [10] Technology roadmap - wind energy, Technical Report, International Energy Agency, 2013. Available. [Online]: http://www.energie-nachrichten.info/file/News/2013/2013-10/Wind_2013_Roadmap.pdf.
- [11] Technology roadmap - solar photovoltaic energy, Technical Report, Energy Technology Perspectives, 2014 edition, International Energy Agency.
- [12] H.F. Gharibeh, A.S. Yazdankhah, M.R. Azizian. Energy management of fuel cell electric vehicles based on working condition identification of energy storage systems, vehicle driving performance, and dynamic power factor, Journal of Energy Storage, vol. 31, Oct. 2020.
- [13] D. Li, A. Zouma, J.T. Liao, H.T. Yang. An energy management strategy with renewable energy and energy storage system for a large electric vehicle charging station, eTransportation, vol. 6, Nov. 2020.
- [14] R. Machlev, N. Zargari, N.R. Chowdhury, J. Belikov, Y. Levron. A review of optimal control methods for energy storage systems - energy trading, energy balancing and electric vehicles, Journal of Energy Storage, vol. 32, Dec. 2020.
- [15] W. Wu, B. Lin, C. Xie, R.J.R. Elliott, J. Radcliffe. Does energy storage provide a profitable second life for electric vehicle batteries?, Energy Economics, vol. 92, Oct. 2020.
- [16] H.S. Salama, I. Vokony. Comparison of different electric vehicle integration approaches in presence of photovoltaic and superconducting magnetic energy storage systems, Journal of Cleaner Production, vol. 260, Jul. 2020.
- [17] T. Zhu, R. Lot, R.G.A. Wills, X. Yan. Sizing a battery-supercapacitor energy storage system with battery degradation consideration for high-performance electric vehicles, Energy, vol. 208, Oct. 2020.
- [18] J. Liu, C. Zhong. An economic evaluation of the coordination between electric vehicle storage and distributed renewable energy, Energy, vol. 186, Nov. 2019.
- [19] J. Adamec, M. Danko, M. Taraba, P. Drgona. Analysis of selected energy storage for electric vehicle on the lithium based, Transportation Research Procedia, vol. 40, pp. 127-131, 2019.
- [20] N. Vukajlović, D. Milićević, B. Dumnić, B. Popadić. Comparative analysis of the supercapacitor influence on lithium battery cycle life in electric vehicle energy storage, Journal of Energy Storage, vol. 31, Oct. 2020.
- [21] M.A. Hannan, M.M. Hoque, A. Mohamed, A. Ayob. Review of energy storage systems for electric vehicle applications: Issues and challenges, Renewable and Sustainable Energy Reviews, vol. 69, pp. 771-789, Mar. 2017.
- [22] I.S. Bayram, S. Galloway, G. Burt. A probabilistic capacity planning methodology for plug-in electric vehicle charging lots with on-site energy storage systems, Journal of Energy Storage, vol. 32, Dec. 2020.
- [23] J. Hu, D. Liu, C. Du, F. Yan, C. Lv. Intelligent energy management strategy of hybrid energy storage system for electric vehicle based on driving pattern recognition, Energy, vol. 198, May 2020.
- [24] S.R. Salkuti, "Electrochemical batteries for smart grid applications", International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 3, pp. 1849-1856, Jun. 2021.
- [25] S.R. Salkuti, C.M. Jung, "Comparative analysis of storage techniques for a grid with renewable energy sources", International Journal of Engineering & Technology, vol. 7, no. 3, pp. 970-976, 2018.
- [26] J. Jin, Y. Xu, Z. Yang. Optimal deadline scheduling for electric vehicle charging with energy storage and random supply, Automatica, vol. 11, Sept. 2020.
- [27] S.R. Salkuti, C.M. Jung, "Overview of Energy Storage Technologies: A Techno-Economic Comparison", International Journal of Applied Engineering Research, vol. 12, no. 22, pp. 12872-12879, Nov. 2017.
- [28] S.R. Salkuti, "Comparative analysis of electrochemical energy storage technologies for smart grid", TELKOMNIKA Telecommunication, Computing, Electronics and Control, vol. 18, no. 4, pp. 2118-2124, Aug. 2020.
- [29] T. Yi, X. Cheng, Y. Chen, J. Liu. Joint optimization of charging station and energy storage economic capacity based on the effect of alternative energy storage of electric vehicle, Energy, vol. 208, Oct. 2020.
- [30] L. Haupt, M. Schöpf, L. Wederhake, M. Weibelzahl. The influence of electric vehicle charging strategies on the sizing of electrical energy storage systems in charging hub microgrids, Applied Energy, vol. 273, Sept. 2020.
- [31] S.R. Salkuti, "Large scale electricity storage technology options for smart grid", International Journal of Engineering & Technology, vol. 7, no. 2, pp. 635-639, Apr. 2018.
- [32] S.R. Salkuti, "Energy Storage Technologies for Smart Grid: A Comprehensive Review", Majlesi Journal of Electrical Engineering, Vol. 14, No. 1, pp. 39-48, Mar. 2020.
- [33] P. Chinnasa, W. Ponhan, W. Choawunklang. Modeling and simulation of a LaCoO₃ Nanofibers /CNT electrode for supercapacitor application, Journal of Physics: Conference Series, pp. 1-5, 2019.
- [34] M. Khalid. A Review on the Selected Applications of Battery-Supercapacitor Hybrid Energy Storage Systems for Microgrids, Energies, vol. 12, no. 23, 2019.
- [35] M. Kroupa, G.J. Offer, J. Kosek. Modelling of Supercapacitors: Factors Influencing Performance, Journal of The Electrochemical Society, vol. 163, no. 10, pp. A2475-A2487, 2016.
- [36] H.J. Jabir, J. The, D. Ishak, H. Abunima. Impacts of Demand-Side Management on Electrical Power Systems: A Review, Energies, vol. 11, no. 5, 2018.
- [37] A. Mohammad, R. Zamora, T.T. Lie. Integration of Electric Vehicles in the Distribution Network: A Review of PV Based Electric Vehicle Modelling, Energies, vol. 13, no. 17, 2020.
- [38] C.T. Ma. System Planning of Grid-Connected Electric Vehicle Charging Stations and Key Technologies: A Review, Energies, vol. 12, no. 21, 2019.
- [39] A. Simpson. Cost-Benefit Analysis of Plug-In Hybrid Electric Vehicle Technology, National Renewable Energy Laboratory, Conference Paper, NREL/CP-540-40485, Nov. 2006.

Detecting Unauthorized Network Intrusion based on Network Traffic using Behavior Analysis Techniques

Nguyen Tung Lam
Information Assurance Department
FPT University, Hanoi
Vietnam

Abstract—Nowadays, network intrusion detection is an essential problem because cyber-attacks are increasing in both the number and extent of the danger. Network intrusion techniques often use various methods to bypass the oversight of anomaly detection and surveillance systems. This paper proposes to use behavior analysis techniques, machine learning, and deep learning algorithms for the task of detecting network intrusions. The practical and scientific significance of our paper includes two issues: (1) Regarding the process of selecting and extracting features: instead of using typical abnormal behaviors of attacks, this study will use statistical behaviors that are easy to calculate and extract while still ensuring the effectiveness of the method; (2) Regarding the detection process, this study proposes to use the Random Forest (RF) classification algorithm, the Multilayer Perceptron (MLP) and the Convolutional Neural Network (CNN) deep learning model. The experimental results in Section IV have proven that our proposal in this paper is completely correct and reasonable. Based on the results shown in Section IV, this study has provided network surveillance systems with a number of abnormal behaviors as the basis for detecting network intrusions.

Keywords—Network intrusion detection; abnormal behaviors; IDS 2018 dataset; deep learning and machine learning

I. INTRODUCTION

Unauthorized intrusion techniques are a dangerous attack form, have been growing rapidly in both the number of recorded attacks and the extent of damage that it causes to organizations or enterprises. Therefore, the task of early detecting and warning signs of cyber-attack campaigns is essential nowadays. Currently, there are two main methods to detect network intrusions: signature-based method through rulesets and anomaly-based method based on analyzing data and statistics to seek abnormal characteristics in the network [1], [2], [3]. The signature-based method has the ability to detect network intrusions quickly and accurately, but it is not possible to detect new attack techniques [1]. The anomaly-based method not only has the ability to detect attacks but also has the ability to detect abnormal behaviors, but it requires complex computation and processing processes and its accuracy is not high. The anomaly-based method is often based on two main techniques to classify abnormal and normal behavior, machine learning and deep learning [1], [2]. So clearly, regarding the network intrusion detection method using machine learning or deep learning, the most important factor is how to identify normal behavior and abnormal behavior. The studies [4, 5] focused on extracting abnormal characteristics and behaviors based on specific attack techniques. However,

we noticed that such an approach could quickly and accurately detect attacks based on specific datasets, but when using other datasets, it is difficult to detect cyber-attack techniques. Therefore, this paper proposes a new network intrusion detection method using deep learning and machine learning algorithms including RF, MLP, CNN based on analyzing behaviors of network traffic. Accordingly, in this paper, we will not find ways to analyze abnormal behavior in network data, we only try to statistic the behavior of network traffic and then use machine learning and deep learning algorithms for analysis and evaluation. With this approach, this study will reduce many steps in finding and extracting abnormal behavior of network intrusion techniques. For the experimental dataset, PCAP files in the IDS 2018 dataset [6] will be selected and used. The study [7] listed and analyzed a number of datasets typically used for detecting cyber-attacks such as DARPA/KDD Cup99, CAIDA, NSL-KDD, ISCX 2012, UNSW-NB15, IDS 2018, etc. In which, the IDS 2018 dataset is built and developed in accordance with real network systems. Therefore, this study will use the IDS 2018 dataset to conduct experiments of cyber-attack detection methods.

II. RELATED WORKS

In the study [8], Vikash Kumar et al. proposed a cyber-attack classification method using UNSW-NB15 and rulesets. Nour Moustafa et al. [9] proposed Geometric Area Analysis Technique for cyber-attack detection using Trapezoidal Area Estimation. This study used UNSW-NB15 and NSL-KDD datasets to conduct experiments in order to evaluate the effectiveness of the proposed method. The experimental results in this study showed the superiority of the UNSW-NB15 dataset compared to the NSL-KDD dataset.

In addition, the study [10] presented a scalable framework for building an effective and lightweight anomaly detection system based on two well-known datasets, the NSL-KDD and UNSW-NB15.

Sikha Bagui et al. proposed in their study [11] a method to detect cyber-attacks based on the Naïve Bayes and Decision Tree (J48) machine learning algorithms. The team [11] used these two algorithms in turn for classifying components of cyber-attacks in the UNSW-NB15 dataset.

The study [12] proposed a cyber-attack detection model using the stacking technique. In their model, the training process uses some machine learning algorithms including K-nearest Neighbor (KNN), Decision Tree (DT) and Logistic

Regression (LR) to build the model based on the UNSW-NB15 and UGR'16 datasets.

The study [13] performed an evaluation of the efficiency of 8 machine learning algorithms (2-layers and 3-layers) for network intrusion detection.

The study [14] presented a DDOS attack detection method using a comprehensive simulation technique of DDOS attacks.

In the study [15], Cho et al. proposed two tasks: detecting cyber-attacks using machine learning algorithms and optimizing features using algorithms such as IG, PCA. Experimental results showed that the team's proposals were relatively good. However, because feature optimization algorithms have large computational times and high complexity, a large calculation system is required. In addition, Cho et al. [16, 17, 24, 25] proposed a method to detect cyber-attacks based on network traffic using machine learning and deep learning algorithms.

In the study [18], Zhao et al. proposed a botnet detection method based on analyzing abnormal behaviors of traffic and flow. Besides, the approach to detect botnet and cyber-attack using the CTU 13 dataset was proposed by Chowdhury et al. [19]. In addition, Ahmed [20] proposed using the ANN deep learning algorithm to classify abnormal connections.

III. NETWORK INTRUSION DETECTION METHOD USING BEHAVIOR ANALYSIS TECHNIQUES

The facts show that with the approach of detecting unauthorized network intrusion using behavior analysis techniques, systems need to perform two main tasks: i) defining abnormal behavior. This definition process is the task of selecting and extracting features, ii) a method of classifying behaviors. This process uses machine learning or deep learning algorithms to classify the behaviors that have just been built in the task (i). We will delve into analyzing and clarifying this content in the next section of the paper.

A. Selecting and Extracting Features

This paper uses the CICFlowMenter tool [21] to handle network traffic. This tool has a function analyzing network traffic into 76 features [16, 17]. These features were presented in detail in the studies [17, 24].

B. Detection Method

As mentioned above, in order to classify intrusion behavior in network traffic, this paper uses a combination of machine learning and deep learning algorithms including Random Forest, CNN, and MLP. These algorithms are being studied and applied in many different problems of the recognition field.

In this, the Random Forest algorithm is a supervised machine learning algorithm researched and developed by [22]. The studies [1, 16] have shown that this algorithm is currently the best classification algorithm because it has a simple operation principle, is easy to calculate and install, especially has low calculation and classification time. The study [22] presented the operating principle and the mathematical model of this algorithm in detail. This paper will use the Random Forest algorithm with standard parameters. We only change the

number of random trees in the algorithm to find and conclude the best model of the algorithm with this experimental dataset.

Regarding the MLP network, the study [23] presented in detail the architecture of an MLP network that is built by simulating the way neurons work in the human brain. MLP networks usually have 3 or more layers including 1 input layer, 1 output layer, and more than 1 hidden layer. Besides, the efficiency of the MLP network depends on the activation function. In this paper, we will tune activation functions to evaluate the effectiveness and suitability of activation functions for the network intrusion detection task.

Finally, the CNN network is defined as a set of basic layers including convolution layer + nonlinear layer, fully connected layer. The detailed structure of CNN as well as the terms: stride, padding, MaxPooling are presented in detail in the paper [23]. In which, the ReLU activation function is used.

IV. EXPERIMENTS AND EVALUATION

A. Experimental Dataset and Scenarios

The experimental dataset is extracted from IDS 2018 Dataset with three types of attacks: Bot (Botnet), Dos, and HTTP-attacks. This dataset is divided into 2 sub-datasets with a total of 762,000 records. In which: the first sub-dataset has two labels: 0 (Benign - clean) and 1 (Bot - malicious); the second sub-dataset has three labels: 0 (Benign - clean), 1 (Dos - malicious), 2 (HTTP-attacks - malicious). We use 70% of this dataset for training and the remaining 30% for testing. Besides, in this paper, to see the effectiveness of the proposed method, we will proceed to refine the parameters of each algorithm to find the most optimal model and architecture.

B. Measures to Evaluate the Results of the Algorithm

The following measures will be used in this paper to evaluate the accuracy of models:

- Accuracy: The ratio between the number of samples classified correctly and total number of samples. Accuracy is calculated by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where: TP - True positive: The number of malicious samples classified correctly; FN - False negative: The number of malicious samples classified as normal; TN - True negative: The number of normal samples classified correctly; FP - False positive: The number of normal samples classified as malicious.

- Recall: is the ratio of true positive points to the total number of real positive points (TP + FN). A high recall means that the TP is high and the rate of missing really positive points is low.

$$Recall = \frac{TP}{TP + FN}$$

- Precision: is the ratio of true positive points to the total number of points classified as positive (TP + FP).

$$Precision = \frac{TP}{TP + FP}$$

- F1-score: is harmonic mean of precision and recall. The higher the F1, the better the classifier.

$$F1 = \frac{2 \times \text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}}$$

C. Experimental Results

1) Experimental results with random forest

a) 2-classes dataset: Table I lists the experimental results of network intrusion detection applying the Random Forest algorithm with 10, 50, 100 trees using the 2-labels dataset.

TABLE I. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING RANDOM FOREST ALGORITHM WITH THE 2-LABELS DATASET

The number of trees	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
10	99.967654	99.967657	99.967679	99.967654
50	99.995733	99.995733	99.995734	99.995733
100	99.988050	99.988050	99.988050	99.988050

From Table I, could see that the algorithm has the highest Accuracy and Precision (99.996%) when the number of decision trees is 50. Besides, when the number of decision trees is changed from 10 to 100, the accuracy of the algorithm does not change much. This shows that with the dataset balanced on the ratio of normal and abnormal records, the Random Forest algorithm brings good and stable detection results. Fig. 1 below presents the evaluation results of the confusion matrix when the number of decision trees is 50. From Fig. 1, seeing that the normal and abnormal prediction models all have high accuracy.

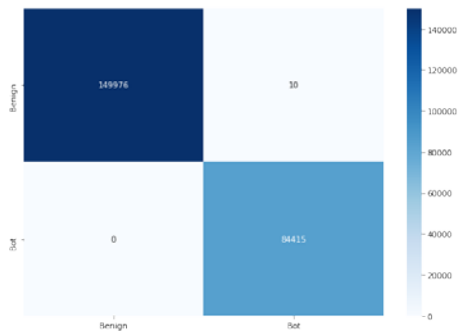


Fig. 1. Confusion Matrix of Random Forest with 50 Trees.

b) 3-classes dataset: Table II lists the experimental results with the 3-labels dataset.

TABLE II. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING RANDOM FOREST ALGORITHM WITH THE 3-LABELS DATASET

The number of trees	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)
10	99.864486	99.837309	99.854480	99.864486
50	99.878638	99.866030	99.868550	99.878638
100	99.886005	99.873035	99.875924	99.886005

Based on the experimental results in Table II, we found that: similar to the 2-labels, the scores obtained with the 3-labels dataset had high results (all over 99%). The Random Forest algorithm gave the best classification results with the number of trees of 100. Comparing the results in Table I and Table II shows that the Random Forest algorithm gave higher efficiency on all measures when using the 2-labels dataset. Confusion Matrix with 100 trees is shown in Fig. 2.

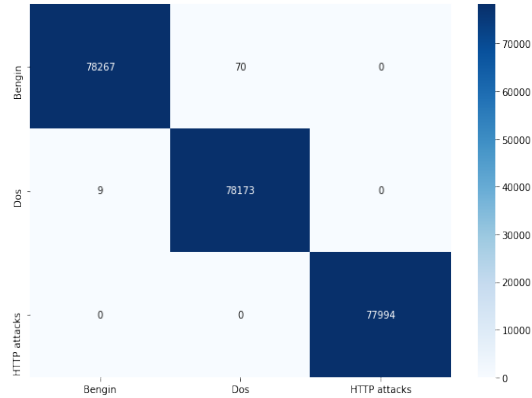


Fig. 2. Confusion Matrix of Random Forest with 100 Trees.

2) Experimental results with MLP

a) 2-classes dataset: From the results shown in Table III, seeing that the MLP model gave very different results when using different activation functions and the number of layers. In particular, with 2 layers, the MLP model gave the best result with ReLU activation. However, when increasing the number of layers to 4, the MLP model had the best results with Logistic activation. But considering accurately detecting the intrusion techniques, the MLP model with ReLU activation still gave a completely better result (reaching 100%). Fig. 3 below is the result of Confusion Matrix when using the ReLU activation function.

From Fig. 3, it can be seen that the MLP model gave prediction results with very high accuracy, with only 32 incorrectly classified records. With this result, it is clear that the MLP model is completely consistent with the purposes and requirements.

TABLE III. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING MLP ALGORITHM WITH THE 2-LABELS DATASET

Parameters		Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Hidden Units	Activation function				
2	identity	94.34	94.51	89.53	91.95
	logistic	97.37	93.65	99.47	96.47
	tanh	81.33	98.33	49.17	65.56
	ReLU	99.90	99.76	99.96	99.86
4	identity	99.06	98.12	99.29	98.71
	logistic	99.62	99.64	99.64	99.48
	tanh	98.18	95.69	99.45	97.53
	ReLU	63.86	63.86	100	77.94

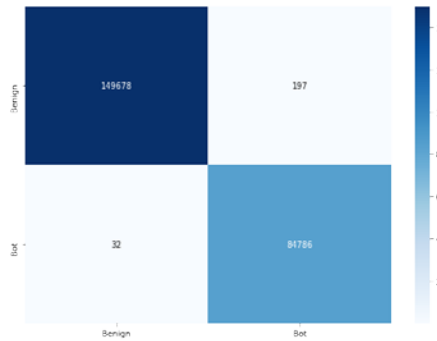


Fig. 3. Confusion Matrix MLP with ReLU Activation Function.

b) 3-classes dataset: The results shown in Table IV show that when increasing the number of classes of the dataset that need to be classified to 3, the F1-score decreased greatly. The average F1-score when using 4 hidden units is higher than when using 2 hidden units. However, the highest F1 was achieved in the case of using 2 hidden units with the Identity activation function. The result of using 4 hidden units and Relu activation function was exceptionally low at 16.67%. Fig. 4 depicts the results of the Confusion Matrix.

TABLE IV. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING MLP ALGORITHM WITH THE 3-LABELS DATASET

Parameters		Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Hidden Units	Activation function				
2	identity	96.10	96.14	96.10	96.10
	logistic	76.77	85.45	76.69	73.31
	tanh	75.35	80.96	75.27	72.14
	ReLU	33.35	11.11	33.33	16.67
4	identity	90.01	90.43	90.01	89.85
	logistic	90.28	90.23	90.26	90.17
	tanh	88.20	89.57	88.17	88.02
	ReLU	33.35	11.11	33.33	16.67

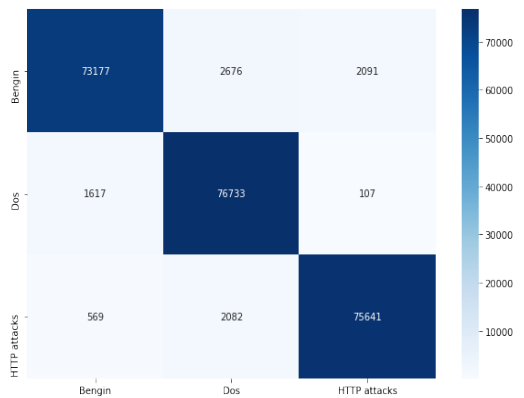


Fig. 4. Confusion Matrix of MLP with Activation Function as Identity.

3) Experimental results with CNN: The CNN network consists of an input layer, hidden layers, and an output layer with corresponding parameters. After many experiments, we found that processing data with Convolution Layers with parameters {filter = 32, 39, 64; filter size = 5; batch size = 32} is optimal. Learning rate parameters of 0.01, 0.001, and 0.0001 were also run to select the most optimal parameter. Based on these results, seeing that a learning rate of 0.0001 gave the best results. Table V describes information about the network models that were selected and tested.

Based on the parameters in Table V, this paper performed with 50 epochs and all Convolution layers used the ReLU activation function.

a) 2-classes dataset: Through results in Table VI, seeing that the CNN model with 1D-CNN achieved very good performance in terms of accuracy, precision, recall, and F1-score. The 1D-CNN 2-layers had the highest performance in 3 models and did not need enough 50 epochs to produce high results. Besides, Fig. 5 presents the accuracy of the training and test process of 1D-CNN 2-layers. Based on it, seeing that this model had an accuracy of approximately 100% after only 23 epochs and maintained that state until the end of the training process. This model detected most attacks (only 8 attack records were not detected). For normal network traffic, the number of false positive record is just 1.

TABLE V. CONFIGURE PARAMETERS OF THE CNN MODEL

Model	Architecture detail
1D-CNN 1 layer	Conv1D(32,5)-MaxPool(2)-FC()-FC()
1D-CNN 2 layers	Conv1D(32,5)-Conv1D(64,5)-MaxPool(2)-FC()-FC()
1D-CNN 3 layers	Conv1D(32,5)-Conv1D(64,5)-MaxPool(2)-Conv1D(39,5)-MaxPool(2)-FC()-FC()

TABLE VI. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING CNN ALGORITHM WITH THE 2-LABELS DATASET

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1D-CNN 1 layer	99.98	99.98	99.98	99.98
1D-CNN 2 layers	99.994	99.994	99.994	99.994
1D-CNN 3 layers	99.9936	99.9936	99.9936	99.9936

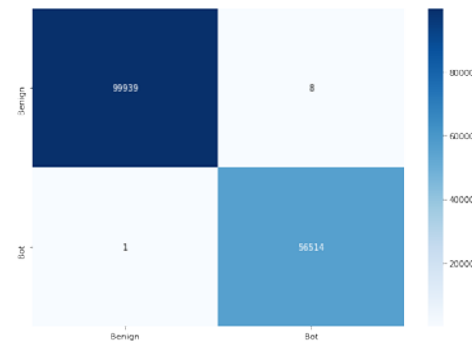


Fig. 5. Confusion Matrix of CNN with 1D-CNN 2 Layers.

b) 3-classes dataset: From Table VII, seeing that the CNN model with 1D-CNN yielded outstanding results on all metrics including accuracy, precision, recall, and F1-score. Besides, for the 3-labels attacker dataset, the 1D-CNN 3-layers had the best performance in the 3 models. The accuracy of the training and testing process of 1D-CNN 3-layers shows in the figure below. It can be seen that this model had an accuracy of approximately 100% after 50 epochs. Fig. 6 below depicts the results of the CNN model with 1D-CNN 3-layers.

TABLE VII. EXPERIMENTAL RESULTS OF NETWORK INTRUSION DETECTION USING CNN ALGORITHM WITH THE 3-LABELS DATASET

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1D-CNN 1 layer	99.937	99.937	99.937	99.937
1D-CNN 2 layers	99.984	99.984	99.984	99.984
1D-CNN 3 layers	99.986	99.986	99.986	99.986

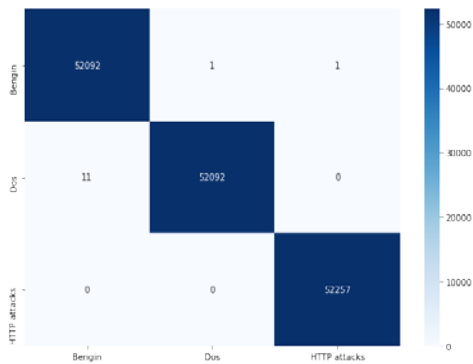


Fig. 6. Confusion Matrix of CNN with 1D-CNN 3 Layers.

D. General Evaluation

Table VIII below shows the overall comparison results of the RF, MLP, CNN classification algorithms with 2-classes and 3-classes dataset.

TABLE VIII. COMPARISON RESULTS OF RANDOM FOREST (RF), MLP, CNN CLASSIFICATION ALGORITHMS

Number of classes	Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
2	RF (50 trees)	99.995	99.995	99.995	99.995
	MLP (relu, 2 units)	99.90	99.76	99.96	99.86
	CNN (1D-CNN 2 layers)	99.994	99.994	99.994	99.994
3	RF (100 trees)	99.967	99.967	99.967	99.967
	MLP (identity, 2 units)	96.10	96.14	96.10	96.10
	CNN (1D-CNN 3 layers)	99.986	99.986	99.986	99.986

With the results shown in Table VIII, with 2-labels and 3-labels dataset, algorithms CNN, Random Forest, and MLP all gave classification results with not too large differences on evaluation metrics. However, in the case of 2 classes, the Random Forest algorithm with 50 trees gave a score about 0.01% higher than CNN (1D-CNN 2-layers). And in the case of 3 classes, the CNN (1D-CNN 3-layers) algorithm gave better classification results than the Random Forest with 100 trees (0.0189% higher). This is not a large number. However, with the actual dataset, it is a quite far distance and has a great impact on the prediction. Therefore, depending on the model of the problem, we will build according to the Random Forest or CNN algorithms. From the data, seeing that with a large amount of data, the number of incorrectly predicted records of the two algorithms is quite much different. Therefore we recommend using CNN rather than Random Forest or MLP algorithms although we must define the network's architecture including the number of layers, decision function, etc.

V. CONCLUSION

Unauthorized network intrusion techniques will transform increasingly to bypass the surveillance of attack detection systems. This requires intrusion detection systems to be constantly updated on the abnormal signs and behavior of network attacks. In this paper, based on analyzing behaviors of network intrusion in network traffic, we have succeeded in determining attack behaviors and normal behaviors of the network data. The scientific and practical significance of the paper is shown in the classification and feature extraction. Accordingly, in our research, we did not extract typical features of cyber-attacks. Instead, we tried to enumerate fully their components and characteristics in the network and then use machine learning and deep learning algorithms to classify. With this approach, we have greatly reduced the time cost of finding and extracting features of network attacks. In addition, based on the experimental results, we have proven that our approach and proposal in this paper are correct and reasonable. This result shows that the proposal using behavior analysis techniques of network traffic using machine learning and deep learning techniques not only helps to accurately detect network intrusion techniques but also contributes to improving the time of seeking and extracting features. Besides, based on the experimental results of Random Forest, CNN, and MLP algorithms with different parameters, seeing that the 2-label dataset gave better results than the 3-label dataset. This shows that: the more optimal the standardization of models and data is, the more accurate the classification is; should not clearly distinguish the labels of network intrusion techniques in the dataset. In the future, we will research and use other analysis methods to improve the efficiency of the detection method based on this dataset. In particular, because our behavior analysis technique has extracted statistical features of network traffic, these features express the correlation not only in terms of data but also in terms of time. Therefore, it is necessary to have algorithms and analysis methods to highlight the time factor in behavior.

REFERENCES

[1] Gilberto Fernandes Jr., Joel J. P. C. Rodrigues, Luiz Fernando Carvalho, Jalal F. Al-Muhtadi & Mario Lemes Proença Jr., "A comprehensive

- survey on network anomaly detection,” *Telecommunication Systems*, vol. 70, pp. 447–489, 2019.
- [2] Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu, “A survey of network anomaly detection techniques,” *Journal of Network and Computer Applications*, vol. 60, pp 19-31, 2016.
- [3] Kh. Ansam. et al., “Survey of intrusion detection systems: techniques, datasets and challenges,” *Cybersecurity*, vol. 20, pp. 2-20, 2019.
- [4] Sebastián García, Alejandro Zunino, Marcelo Campo, “Survey on network-based botnet detection methods,” *Security Comm. Networks*, 2013. <https://doi.org/10.1002/sec.800>.
- [5] Monowar H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, “Network Anomaly Detection: Methods, Systems and Tools,” *IEEE Communications Surveys & Tutorials*, vol. 16 (1), pp. 303–336, 2014.
- [6] CSE-CIC-IDS2018 on AWS. <https://www.unb.ca/cic/datasets/ids-2018.html>.
- [7] R. Markus., et al., “A survey of network-based intrusion detection data sets,” *Computers & security*, vol. 8, no. 6, pp. 147–167, 2019.
- [8] K. Vikash., et al., “An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset,” *Cluster Computing*, vol. 22, doi: 10.1007/s10586-019-03008-x, 2019.
- [9] N. Moustafa., et al., “Novel Geometric Area Analysis Technique for Anomaly Detection using Trapezoidal Area Estimation on Large-scale Networks,” *IEEE Transactions on Big Data*, vol. 5, no. 4, pp. 2332-7790, 2017.
- [10] N. Moustafa et al., “Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models,” doi: 10.1007/978-3-319-59439-2_5, 2017.
- [11] S. Bagui, et al., “Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset,” *Security and Privacy*, doi: 10.1002/spy2.91, 2019.
- [12] S. Rajagopal., et al., “A predictive model for network intrusion detection using stacking approach,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2734-2741, June 2020.
- [13] S. Rajagopal, et al., “Performance analysis of binary and multiclass models using azure machine learning,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 978-986. February 2020.
- [14] H. H. Ibrahim, et al., “A comprehensive study of distributed Denial-of-Service attack with the detection techniques,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 3685-3694, August 2020.
- [15] Cho Do Xuan, Hoang Thanh, Nguyen Tung Lam, “Optimization of network traffic anomaly detection using machine learning,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2360-2370, 2021.
- [16] Cho Do Xuan, Lai Van Duong, Tisenko Victor Nikolaevich, “Detecting C&C Server in the APT Attack based on Network Traffic using Machine Learning,” *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 11(5), 2020. <http://dx.doi.org/10.14569/IJACSA.2020.0110504>.
- [17] Cho Do Xuan, Hoang Mai Dao, Hoa Dinh 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, 2019.
- [18] David Zhao, Issa Traore, Bassam Sayed, Wei Lu, Sherif Saad, Ali Ghorbani, Dan Garant, “Botnet detection based on traffic behavior analysis and flow intervals,” *Computers & Security*, vol. 39, pp. 2-16, 2013.
- [19] Sudipta Chowdhury, Mojtaba Khanzadeh, Ravi Akula, Fangyan Zhang, Song Zhang, Hugh Medal, Mohammad Marufuzzaman & Linkan Bian, “Botnet detection using graph-based feature clustering,” *Journal of Big Data*, vol. 4, no. 14, 2017.
- [20] Abdulghani Ali Ahmed, Waheb A. Jabbar, Ali Safaa Sadiq, Hiran Patel, *Journal of Ambient Intelligence and Humanized Computing*, 2020. <https://doi.org/10.1007/s12652-020-01848-9>.
- [21] CICFlowMeter. <http://www.netflowmeter.ca/netflowmeter.html>. Accessed 1 November 2020.
- [22] L. Breiman., “Understanding Random Forests: From Theory to Practice,” *Machine Learning*, vol. 80, no. 1, pp. 5-32, 2017.
- [23] Samaneh MahdaviFar, Ali A. Ghorbani, “Application of deep learning to cybersecurity: A survey,” *Neurocomputing*, vol. 347, pp. 149–176, 2019.
- [24] Cho Do Xuan, Dao M.H, “A novel approach for APT attack detection based on combined deep learning model,” *Neural Comput & Applic*, 2021. <https://doi.org/10.1007/s00521-021-05952-5>.
- [25] Cho Do Xuan, “Detecting APT Attacks Based on Network Traffic Using Machine Learning,” *Journal of Web Engineering*, vol. 20(1), pp. 171-190, 2021.

An Integrated Implementation Framework for an Efficient Transformation to Online Education

Ahmed Al-Hunaiyyan¹, Salah Al-Sharhan², Rana Alhajri³, Andrew Bimba⁴

Computers and Info. Systems Department, Public Authority for Applied Education and Training (PAAET), Kuwait¹

School of Business and Information Technology, College of North Atlantic- Qatar²

Computer Department, Public Authority for Applied Education and Training (PAAET), Kuwait³

College of Information Technology, University Malaya- Malaysia⁴

Abstract—The least developed countries have been tasked with introducing effective e-learning frameworks as they look to overcome technology inadequacies and lack of research support or vision. Ongoing efforts are reliant upon a mixed-methods approach. A systematic literature analysis and a quantitative examination have been undertaken to achieve a thorough assessment of the available data taken from educational facilities in Kuwait. Results show clear support for embracing e-learning, with most participants recognizing its positives when faced with the scope of challenges its practice may incorporate. Consequently, the authors recommend a framework that is integrated to support a smooth upgrade to online teaching in a manner that furthers the efficacy and understanding of e-learning potential in the context of education in Kuwait and neighboring countries, with a particular focus on how to function during a pandemic lockdown. The proposed framework is structured according to five key tiers: infrastructure, e-learning delivery, LMS, e-Content, and user portal. In support of this, a model of e-Content development is proposed to assist with the establishment and execution of educational materials, in particular, to cope with the lack of digital learning materials in Arabic.

Keywords—Distance learning; e-learning framework; COVID-19 pandemic; e-learning delivery; e-Content

I. INTRODUCTION

The worldwide technology boom has resulted in the transfer of education via desktops to learn remotely, meaning the increased potential for fresh and innovative education approaches. Technological progress supports educational facilities having the capacity to adapt to remote teaching. Still, it would be false to claim that all educators have seamlessly achieved this transformation because of a lack of direction or advice on how it should best be achieved. Indeed, despite its advantages, turning to e-learning at a challenging time – such as during the pandemic lockdown – can prove complicated and damaging, involving numerous connected factors that need to be analyzed and adapted to meet robust models and frameworks. Such e-learning frameworks and models need to ensure this new generation of learners are effectively connected, enhancing their ability to learn in isolation and take advantage of new software [1, 2]. Those decision-makers involved in establishing a new learning landscape are focused on organizing educational facilities and cities into exemplars of learning development that is in balance with each other [3]. As more moves are being made to establish Smart Cities that

direct job markets and businesses toward knowledge-based contributions, the OECD is calling for educational institutions to focus on more than education alone. Indeed, they call for universities to be aligned with knowledge-concentrated job markets, allowing those who graduate to fit in smoothly with the smart city vision [4].

Innovative frameworks are being launched, as researchers gain a greater awareness of potential when answering the problems and pitfalls of technology upgrades. In the context of this paper, the definition of ‘framework’ matches that defined by [5]: “standardized set of concepts, practices, and criteria, applicable on a certain type of issue, to provide solving solutions.” Education providers have had little choice to move from teaching in-person to remote solutions due to the pandemic. Therefore, this transformation's speed without a tried and trusted process, advice, or practical execution framework means that the results have not always been positive – particularly among those that remain less developed technologically speaking [6, 7]. The effective rollout of remote and blended learning solutions in less developed countries demands an appreciation of how technology best facilitates both learning and teaching practices that lead to positive outcomes [8].

Most importantly, this transformation requires restructuring each educational facility's dynamics to suit each student's learning capacity and securing a move from traditional teaching influence on digital influence that is just as interactive and impactful as before, if not more so. The potential for making this shift has been fully realized thanks to e-learning versatility, but that does not mean that all education providers worldwide have done so in a robust fashion – particularly in less developed nations [9, 10]. The absence of applicable frameworks for guiding educators in securing this transformation into a new age of education is undoubtedly one of the key factors holding systems back [11, 12, 13, 6, 14].

Kuwait remains a developing nation, meaning that all the benefits of ICTs and e-learning potential were still behind many other countries prior to pandemic lockdown challenges. Most education providers were still focused on traditional methods, therefore, learning via classroom settings [9, 15]. Although there have been numerous moves to facilitate online learning, few have been successful [15, 9]. So, although Kuwait has seen some education providers becoming familiar with online learning potential, as Alkharang [9] confirms, its

progress has not matched that of other nations due to the need to upgrade current systems and technology. Section 4 of this paper looks further into Kuwait to analyze precisely why e-learning adoption has caused certain rifts and dilemmas. The focus is on highlighting key factors to enable identifying a robust framework required for effective integration.

A. Research Objectives

Several e-learning frameworks have been used to solve unique e-learning problems. Nevertheless, these frameworks mainly cover issues of e-learning systems development, application and adoption while the development of electronic content (e-Content) and its role in the educational process is missing especially digital Arabic learning content [8]. Therefore, the aim of the current study is to develop an integrated framework in the context of virtual and smart classrooms environment to effectively adapt the efficient use of technologies in the Kuwaiti national educational systems and the developing countries in the region. The framework is developed in the light of identified critical factors, international practices, and the outcome of the survey study presented in Section 4. Electronic content utilization, which has been identified as vital to an effective e-learning framework, is assessed here as a singular entity rather than as part of a whole.

The shift to online education requires digitizing the conventional curricula and transferring the curricula into interactive online subjects that ensure an efficient learning process and provide pedagogical strategies to achieve the educational objectives and learning outcomes. The importance of this analysis is the ability to successfully merge the technical and pedagogical aspects of education in a manner that maximizes the utilization of e-Content, that makes the most of teachers' roles regardless of how they are able to interact with students, and that integrates with the latest technological innovations [16]. Furthermore, the framework focuses on effectively addressing the challenges specific to Kuwait and similar neighboring countries, incorporating the issues identified by the pandemic lockdown.

The paper begins with a literature review in Section 2, then moves on to the research methodology in Section 3, the case study in Section 4, and the e-learning framework in Section 5. The final section concludes the research and proposes future directions.

II. LITERATURE REVIEW

Developing a robust e-learning system requires a focused examination of the resources and avenues that the education provider has access to. As proposed [17], the design, assessment, and rolling out of new frameworks demand firm insights into the various features and versatilities of online technologies. Plus, this paper explores the technical and pedagogical aspects of teaching remotely, together with e-Content utilization to meet key learning goals.

A. E-Learning Delivery

Various concepts of e-learning and remote education have drawn a wide range of research initiatives. Nevertheless, strong similarities exist to show that all different concepts can

be brought under the banner of interactions between students and educators that occur regardless of time and place [18]. E-learning itself can be defined as the utilization of online technology in a manner that allows for learning remotely [19]. However, various online learning setups are available such as standalone or virtual classroom options, the use of simulations, and embedded or blended learning [20, 21, 22]. Despite this, two distinct forms of online learning have been identified. The first is an all-encompassing approach, demanding every aspect to be carried out in an online or virtual form and no teacher or student interactions occurring in person [23]. As perhaps expected, therefore, the second form involves a blended approach, with online technology utilized as much as possible but not necessarily defining every interaction – more of a 'Smart classroom' that facilitates online learning but is still led by recognized traditional methods [22, 24]. Essential for both these approaches is the adoption of Learning Management Systems (LMSs), via which students and educators can communicate.

Moving education online alters learning structure, establishing new and innovative forms of teaching that can be either synchronous or asynchronous. The asynchronous approach requires the educator to provide solutions in real-time, whereas an asynchronous technique does not rely on real-time interaction [25]. The merging of various learning techniques has been trialed in Poland (Ożadowicz, 2020), tackling pandemic lockdown challenges with a combination of synchronous and asynchronous approaches. Such a strategy was adopted to suit the challenges of engineering, being a subject that requires a certain degree of hands-on work on-site and in the laboratory. Hence, students were still able to perform the practical aspects of their work while also learning from teachers online. Research elsewhere [26, 27], however, claims that teaching in person should not be wholly sidelined and remains the most important aspect of any education, despite the flexibility allowed by online potential. These researchers warn against adopting an asynchronous approach for the sake of convenience alone, stressing the more laid back and reflective benefits of teaching in person. Nevertheless, other researchers seek to merge learning approaches to make the best of all avenues, calling for a versatile and hybrid outlook on education [28, 29]. A wide range of digital learning approaches has been trialed since the pandemic began. Nerantzi (2020), for example, promoted peer instruction and flipped learning to education providers in the UK. Such methods focus on simulating knowledge accrual in merged settings, so long as students are engaged in an active sense. The peer instruction approach involves scaffolding, using peer-peer interaction, which promotes active engagement and promotes conceptual understanding outside the classroom. The proposed technique's main aim is to provide instructors with modules to maximize student engagement and learning.

With more and more studies of online learning potential taking place, the various frameworks and practicalities involved are under constant scrutiny. Examples of key priorities for researchers include designing online teaching materials and their sustainability [30, 31, 32]. Additional areas of concern are the preparedness of teachers, the willingness to embrace technology, the form of delivery [14, 28, 29, 33], the

role of the smart classroom [34], remote learning [35, 1], and mobile learning [36, 37]. Other researchers look to be more specific by stressing Critical Success Factors (CSFs), defined as institutional management, learning environment, instructional design, service support, and course evaluation [38]. Similarly, having a recognized smart classroom framework is recommended by Al-Sharhan [34] as a way of narrowing down the focus on remote learning. This method involves directing the smart classroom capabilities so that curriculum and ICT features become seamlessly integrated. An approach that coincides with [36] put forward a form of remote learning that blends mobile technology with established pedagogy. This alternative involves utilizing mobile technology within a shared virtual learning zone, enhanced to incorporate e-Content and LMS aspects.

Support for effective remote learning involves a range of obstacles and dilemmas. Some studies have analyzed these issues in-depth, exploring problems related to communication, timekeeping, infrastructure, and user familiarity with advancing technology [39]. A varied scope of research concerning how education can be readjusted can be found around the world. For example, in the Philippines, [7] explored the consequences of cutting out face-to-face interaction completely. This led to identifying issues that included teaching with no particular style, poor standards, and the ability to learn solely from domestic surroundings. In contrast, in South Africa, the consequences of remote learning reliance on students from disadvantaged backgrounds have come under scrutiny [6]. This research was carried out with the motivation to find more flexible learning avenues that allow students from a wide range of social situations to continue to access quality education. The conclusions point to a 'digital divide' as preventing learners from disadvantaged backgrounds from making the most of technology in the same way that others can.

B. e-Content Development

Many approaches, models and frameworks exist when designing quality online learning environments. These approaches assist and guide instructional designers through the design and development of e-Content. Learning management systems (LMS) provide an instructor a way in which to create and deliver content, monitor learners' participation, and assess learners' performance. In fact, many institutions may have one or two content-authoring tools supported by the LMS. These fundamental tools allow institutions to form e-Content solutions that embrace a range of teaching approaches, as well as directing current environments to suit a range of social scenarios, learning capabilities, and gender differences [40, 41, 42]. Such a method, designed to merge student engagement with a range of teaching styles and methods, was proposed in a 2016 study [32]. The focus for this study was on online developers and designers rather than the educators themselves. While, to a certain degree, the attributes that learners bring to their challenges will come to define what they can achieve, it is also the case that an efficacious course design enhanced by

innovative technology will facilitate their potential [32]. As far as e-Content is concerned, six key developmental stages are identified as analysis, design, development, testing, implementation, and evaluation [31]. Furthermore, this study identifies pedagogical concerns in how this new content is driven, incorporating a wide range of influencing factors, such as teaching and learning styles, forms of examination, overall strategies and processes, recognized practice, and the ultimate aims of learning content [31]. The educator remains crucial in all of these contexts, effectively facilitating the impact that digital technologies have on students' learning progress [43, 1].

Instructional design (ID) is integral to establishing robust e-Content [44]. For those who practice ID, they require a strategy that is both systematic and repetitive in how problems are addressed. The process involves establishing the instructional objectives, assessing the key goals and learning domains, clarifying the outcomes required, and drawing on relevant test questions to complement the overall strategy. Five examples of developed ID models were placed under analysis by the OECD: gaming, virtual laboratories, collaborative projects, real-time assessments, and skills-based assessments [4]. In contrast, an alternative approach identified four crucial ID dimensions in the form of function, origin, source, and analysis [45]. These defining dimensions were merged to establish ten effective actions that make up a firm methodological ID strategy. Elsewhere, the prioritization of pedagogical objectives despite lockdown complications has led to creating various methods that cut down on time a learner needs (Cahapay, 2020). This approach requires determining the most vital aspects of a curriculum, then relieving teachers from juggling too many aspects as they help students to progress. Both hybrid and fully online learning strategies are possible in this respect. However, less developed countries will likely encounter far more social and cultural issues adjusting to this new outlook.

III. METHODOLOGY

An analysis has been undertaken to address two categories of research literature looking into how online learning is currently being perceived, including the various strategies and e-learning frameworks proposed for e-learning implementation and the utilization of e-Content. The literature that has been written under the current lockdown situation was favored, but not exclusively. Once carried out, survey research followed to draw in key data thanks to a questionnaire constructed to connect with both students and their educators' experiences. In total, 4,024 participants were involved, all sourced from Kuwaiti educational facilities. The data received was then scrutinized via statistical analysis to identify the most impact of taking up and adjusting to online learning challenges. Fig. 1 shows the theoretical framework adopted for this process [46, 47]. The framework requires a definition of the system under scrutiny as the first stage (educational facilities in Kuwait) before its remit moves on to concerned parties in the second stage (students and their educators).

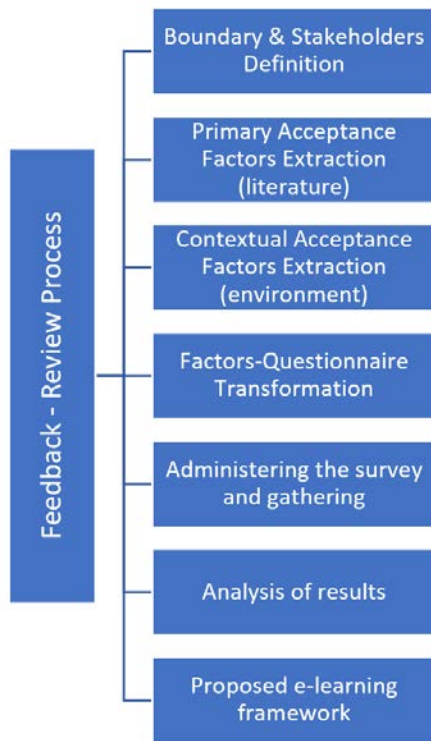


Fig. 1. Theoretical Framework.

The second stage involves utilizing the literature analysis to identify how the challenges of e-learning and its various methods and frameworks are currently being addressed. This assessment highlights the factors needed to analyze the embracing of e-learning. The third stage involves recognizing the contextual issues both teachers and learners face when adjusting to online methods – i.e., the complications of real-life situations. As far as this study is concerned, these contextual issues are gathered from within Kuwait. Once these stages are complete, the factors identified are organized into a table that incorporates the relevant research literature. The fourth stage utilizes the accrued information to form an applicable questionnaire before the survey participants are approached to gather the fifth stage data. An examination of statistical data and recommendations is made in stage six before firm solutions, and frameworks are proposed in stage seven. In support of this work, feedback and review procedures were also in place.

IV. CASE STUDY: EMBRACING ONLINE EDUCATION IN KUWAIT

Reacting to the complications of lockdown and lack of face-to-face teaching, educational providers throughout Kuwait have been debating methods for proceeding with learning objectives, mainly in terms of how online technology can be utilized. All educational facilities were closed due to the pandemic, leading to a wide range of issues and challenges for those at crucial learning and development stages. The survey study explored here is a result of the various dynamics that emerged.

A. Study Sample

The focus is on educators and learners registered with Kuwaiti facilities (from both private and public sectors). Sample responses were gathered within two time periods: first, during a period of pilot study, then as part of the resulting primary survey. Once complete, 4,024 participants had returned their responses, with the demographics incorporated into Table I.

TABLE I. SAMPLE FEATURES

		Frequency (n)	Percent (%)
Gender	Male	864	21.5
	Female	3,160	78.5
Category	Student	2,902	72.1
	Instructor	1,122	27.9
Institution type	Public	2,980	74.1
	Private	1,044	25.9

B. Instrument for Data Collection

The data collection instrument was purpose-designed to suit key issues identified via the research literature [48, 49, 50, 51]. Furthermore, internal issues became apparent in Kuwait's contexture and incorporated [12, 52, 10]. In total, the questionnaire was made up of 13 questions and split into two parts – part one to obtain the demographic features; part two includes ten key questions that could be subjected to a three-point Likert-type scale (1 for disagree, 2 for neutral, and 3 for agree). Once set up, the questionnaire was sent out to contributors via a number of avenues, including email, social media, and e-messaging. A pilot study was carried out so that any issues with the survey questions could be addressed, which resulted in a Cronbach's Alpha score of 0.901, showing that the questions were applicable and could be moved on into the preliminary study stage. However, the pilot study results are included in the overall total, with all results subjected to frequency, percentage, mean, and standard deviation (SD).

C. Results

The preliminary analysis findings are shown in Table II. These results indicate a mean value of 2.10 for item 1, showing that participants were slightly in favor of e-learning. The question of whether e-learning versatility allows respondents to feel confident turned out to be positive – recording a 2.27 mean value. In contrast, an item 3 mean value of 2.35 showed that both learners and educators feel favorable toward online learning for courses of a theoretical context. However, moving on to item 4 shows that this positivity is not reciprocated for learning that depends on laboratory activities – recording a 1.95 mean value. In consideration of synchronous and asynchronous learning objectives, in item 5, synchronous learning received a lower mean value of 2.19 (for the benefits of real-time video conferencing) instead of 2.35 for item 6. These latter two responses show that users prefer the flexibility of learning as and when they see fit instead of having rigorous calendar activities.

TABLE II. LEARNERS' AND EDUCATORS' VIEWS OF ONLINE LEARNING IN RESPONSE TO LOCKDOWN IN KUWAIT

No.	Questions	Disagree		Neutral		Agree		Mean	SD
		Freq.	%	Freq.	%	Freq.	%		
1	Utilizing e-learning is an applicable approach to continuing with studies during lockdown.	1,175	29.2	1,267	31.5	1,582	39.3	2.10	0.822
2	Having access to e-learning means more versatile approaches, allowing students to learn according to the time and place they see fit.	1,010	25.1	928	23.1	2,086	51.8	2.27	0.835
3	E-learning is ideal for theoretical courses.	748	18.6	1,134	28.2	2,142	53.2	2.35	0.774
4	E-learning can replace laboratory sessions as an effective alternative.	1,343	33.4	1,548	38.5	1,133	28.2	1.95	0.783
5	I favor live video chat (synchronous) with both the instructor and other students.	1,233	30.6	788	19.6	2,003	49.8	2.19	0.876
6	I favor analyzing resources and study materials (asynchronous) provided by educators via online platforms (like YouTube).	867	21.5	873	21.7	2,284	56.8	2.35	0.812
7	Effective e-learning means that educators and students need to be trained in its application.	341	8.5	715	17.8	2,968	73.8	2.65	0.629
8	My facility offers adequate training courses for e-learning.	2,063	51.3	990	24.6	971	24.1	1.73	0.825
9	Our infrastructure (hardware, software, networks) is suitable for the challenges of e-learning.	1,268	31.5	1,183	29.4	1,573	39.1	2.08	0.837
10	A range of digital e-learning resources meet the challenges of our curriculum.	1,446	35.9	1,465	36.4	1,113	27.7	1.92	0.793

As item 7 shows, educators and learners both feel that a certain amount of training is essential so they can make the most of e-learning potential (mean 2.65). The feedback indicates that both parties are keen to enhance their understanding of e-learning tools and their capabilities. Compare this to the item 8 result (recording only a 1.73 mean value), and we can conclude that educational facilities are not doing enough of this. Equally as revealing is the 2.08 mean value score for item 9, which shows an agreement with the notion that the current infrastructure is falling short. Additionally, a 1.92 mean value for item 10 indicates that users are not particularly confident in those remote learning applications that are currently being utilized.

The decision to continue the rest of the academic year with online learning sparked concerns and debate between supporters and skeptics. In response, Kuwait's former education minister claimed that e-learning is a preferable solution to having no access at all. In contrast, Kuwait's previous higher education minister criticized the issues related to e-learning rollout, considering that the country has all the infrastructure needed to make this work [53]. This view led to 11 Kuwait University professors signing their agreement that online learning should be utilized for the remainder of the academic year, calling for 'continuity' as the most crucial factor in someone's education [12]. However, there were some opinions to the contrary, with one professional vocal calling to acknowledge that current e-learning capabilities do not meet the curriculum objectives. He focused on ways in which the present capabilities are falling short, both from educators' and learners' perspectives, as well as the accessibility of specific online resources [10]. Concurring, the Kuwaiti Cultural Office's previous head in Washington DC pointed out the many obstacles to effective learning in an online environment. Part of his concerns included how ready both educators and learners are in respect of utilizing a wide range of innovative but unfamiliar resources and an inability to follow and analyze

student progress (Academia 2020). Some of these concerns have been echoed by other academic voices within Kuwait, particularly in the context of readiness [12]. Although facilities worldwide have had no choice but to take this step, that does not mean that the pitfalls should be overlooked. An absence of firm strategies, digital learning resources, and individual readiness are all issues when making such a transformation [54, 7].

V. INTEGRATED IMPLEMENTATION FRAMEWORK

The effective deployment of e-learning resources requires numerous crucial attributes to be in sync, including overall management, technical expertise, and social and cultural dynamics. As a result, the following section looks to weigh up these various issues so that the pathways both educators and learners take can be robust and successful throughout society.

Taking direction from the literature review findings – both the successes and pitfalls encountered – and combined with the questionnaire findings, an effective and integrated implementation process is proposed so that all teaching methods and resources can be merged to result in a productive online learning solution. As online capabilities are so crucial, there is a focus on ensuring that professionals develop these solutions in order to safeguard from any usability issues. Nevertheless, the effectiveness of the proposals depends on a range of issues, including the actions taken by the educational facility itself, the overcoming of technical issues like bandwidth, network security, and productivity, as well as links with the centralized Arabic local and regional data center. An infrastructure that is capable of high performance, therefore, is vital to e-learning success. Consequently, the framework put forward incorporates both local infrastructure and cloud-based architecture. So, the results might rely on either public or private cloud models or a merger of the two depending on the most applicable. Besides, all learners need their applicable mobile device in order to access the necessary resources

irrespective of their location. Should a network be troubled by sluggish performance, this will be inadequate for addressing the teaching and learning challenges involved. Indeed, these issues need to be prioritized within the framework solutions rather than seeking to address them as subsequent performance problems.

As set out in Fig. 2, the e-learning infrastructure advised has been established according to practical issues and contexts particular to Kuwait's challenges, so incorporating both social and cultural factors. The various implementation stages involved are depicted according to specific tiers, all of which are designed to make up an effective distance and blended learning system (DBLS). A bottom-up approach is organized so that implementation can be effectively organized as follows: infrastructure; learning management system; interactive content repository (e-Content); e-learning delivery; and user gateway (portal). Each level involves precise components and design details to inform developers. Plus, an ongoing focus on quality and effectiveness should be pursued according to the content of user reviews and feedback.

The framework incorporates a field of implementation, ranging from knowledge sharing to user collaborations and including platforms for users to engage and feedback. Also, it can be recognized that the various tier levels stand for the Kuwait learning gateway and performance management unit. This tier's key role, alongside quality assurance aspects, is to monitor and improve pedagogical standards to effectively realize learning goals. Doing so is crucial for making sure that teachers, technical advisors, or any managerial figures can monitor the performance of the various tools and resources involved. Quality Assurance (QA) is one of the most critical factors that focus on continuous improvement at several levels that should be applied for each tier and incorporated with every phase of the learning process to ensure its accuracy and effectiveness. The importance of the quality assurance and academic standards describe the level of achievement that the

e-learning process and strategies achieve its broader objectives [55, 56].

A. Tier (1): Infrastructure Tier

The infrastructure tier focuses on realizing quality computing resources that offer excellent user-friendliness thanks to either public cloud access or a hybrid approach. Having a hybrid approach means that features are made available via both public and private cloud resources but have to work together seamlessly. The intention is to produce an innovative and fully integrated infrastructure that does not lag behind in any technological aspects and allows learning forums to be merged and utilized. It should also allow those overseeing the educational facilities to make way for incoming advances regarding cloud computing and new technology. As well as being user-friendly and easily accessed, the infrastructure will offer a high performance that does not hold learning back, with users able to carry out their activities dynamically and interactively while also having excellent file storage. Other key aspects include general usability and automation for utilizing an assortment of storage solutions via a centralized approach. Furthermore, cloud computing needs to be upgraded where new IT features and capabilities emerge. The system remains up to date without involving manual maintenance from staff members within the educational facilities. Lastly, the infrastructure needs to be planned to reduce levels of support and technical costs as much as possible.

B. Tier (2): e-Learning Delivery

An effective distance-learning solution needs to incorporate a wide range of features to offer a robust system, for example, facilitating communication, smart classrooms, and dynamic multimedia resources that are merged with LMS benefits and easily obtainable remotely well as on-site. This e-Learning delivery stage requires a range of instructional channels.

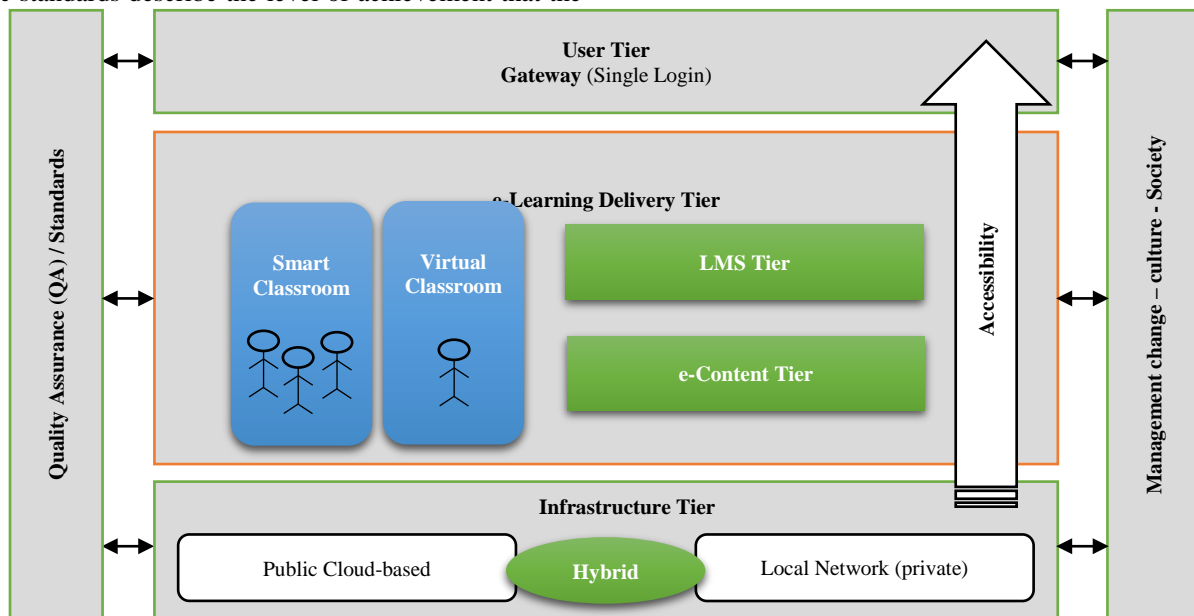


Fig. 2. The Integrated E-Learning Implementation Framework.

Smart Classroom Component (Blended): Typically, the smart classroom's function is to facilitate merged learning processes in a manner that is unique to each learner, incorporating the latest innovations, and yet present within a traditional classroom environment and overseen by teaching professional. In essence, the smart classroom stands for the upgrade of a conventional classroom environment to merge with modern capabilities. The qualities of recognized teaching methods and innovative technologies are combined, therefore, to the benefit of both learners and teaching professionals.

Virtual Classrooms (Distance): When restrictive scenarios like the Covid-19 lockdown emerge, the benefit of a virtual classroom will be seized upon so that learners and educators can continue to meet and pursue their objectives. This type of learning still occurs in real-time rather than as the student feels fit, with Google, MS Teams, and Moodle all being examples of this approach being taken up successfully worldwide. When functioning to their maximum potential, the virtual classroom will be merged with LMS, as per tier 3 below.

C. Tier (3): Learning Management System

This is a crucial tier for realizing potential, though it is sometimes challenging to summarize. Typically, LMS requires a software application to organize the performance, monitoring, and reporting of the educational curriculum [57, 58]. The two tiers of LMS and e-learning delivery combined encompass learning settings that oversee all features, such as course material, content links, student enrollment, and examinations. Educators can monitor how learners perform, which is why there needs to be a focus on both design and integration so that professionals can carry out all their duties while ensuring each student's learning journey is wholly facilitated. Neither location nor time factors should be an issue so long as students have access as and when they please. Simultaneously, educators can make sure all required resources are accessible, and that content is interactive for learning purposes. LMS is central to the system's overall delivery, allowing for course materials to be presented, for students to be monitored, and for performance to be assessed. Plus, a range of different assessment options is incorporated, overcoming one of the key issue's educators face during the lockdown. Therefore, a focus on outcomes and content can remain with LMS acting in place of traditional forms of scrutiny.

D. Tier (4): Interactive e-Content

Among the most crucial aspects of an online learning system, interactive e-Content deserves to be analyzed. Effective design and delivery of educational programs are impossible without effective pedagogical integration, alongside access regardless of location or time of day.

This fourth-tier links those critical elements of tiers 3 and 5, positioned between the process of learning and its outcomes. While possible to overlook, interactive e-Content is essential for any robust e-learning solution because, without sound design and technical development, neither the learning process nor its outcomes can run efficiently. Those working on its development need to be familiar with the creation of Learning Objects (LOs), which might be an obscure scientific

term to many but account for a number of vital interactive elements, including images and audio. Skilled developers will recognize each LO (or SCO [Sharable Content Object]) as containing several key factors:

- Pedagogical aspects are applicable to the learning object.
- Concepts clarified and delivered via multimedia-based approaches.
- Actions and activities linked with the curriculum's key learning goals.
- The ability to examine how students are interacting with the materials and resources.

The accessibility of the digital e-Content range is vital. Linking students with the heart of the curriculum (as per tier 3), ensuring that no insurmountable obstacles stand between their ability to achieve the best possible results, effectively overcoming challenges such as those posed by lockdown thanks to the versatility of distance learning. Educators can keep up to date with how their students are progressing thanks to LMS's insights, which, if utilized effectively, will inform their teaching practice and allow for final examinations (as per tier 5). In establishing effective e-Content, developers are advised to focus on five key aspects: Analysis, Design, Development, Application, and Evaluation – or the ADDIE [59]. A firm analysis process is advised to ensure that the e-Content applied performs according to the specific objectives. This paper's development model is organized according to a 2010 research [17], primarily due to its focus on cultural dynamics. Fig. 3 illustrates the e-Content development model describing the efficient steps requires to develop electronic content based on software engineering practices, including instructional design strategies.

E. Tier (5): User (Portal)

The learning gateway (user tier) concerns the convenience of access that each user (learner or instructor) enjoys – which any system developer will also experience. Effectively, the gateway functions as a top-level view, allowing a single sign-in portal and numerous resources applicable to the learning or teaching processes. The key aspects of a learning gateway incorporate the capacity to oversee e-learning systems. The objectives are to contain and present interactive digital e-Content, to link administrative and educational materials, to allow for correspondence between the various system users, and to effectively represent all parties in a manner that suits them [60]. Whether a professional is active in a teaching capacity, as a supervisor, manager, or indeed as an overseeing policymaker, an effective system will facilitate their activities, which will have a knock-on positive contribution concerning how professionals communicate and make decisions. This tier is also vital for establishing those tools which allow professionals to assess progress, for example, via generated performance reports. Plus, this tier tends to include a centralized dashboard that adheres to the Business Intelligence (BI) tool, which allows all professionals involved to analyze any performance reports that might enhance or inform their methods.

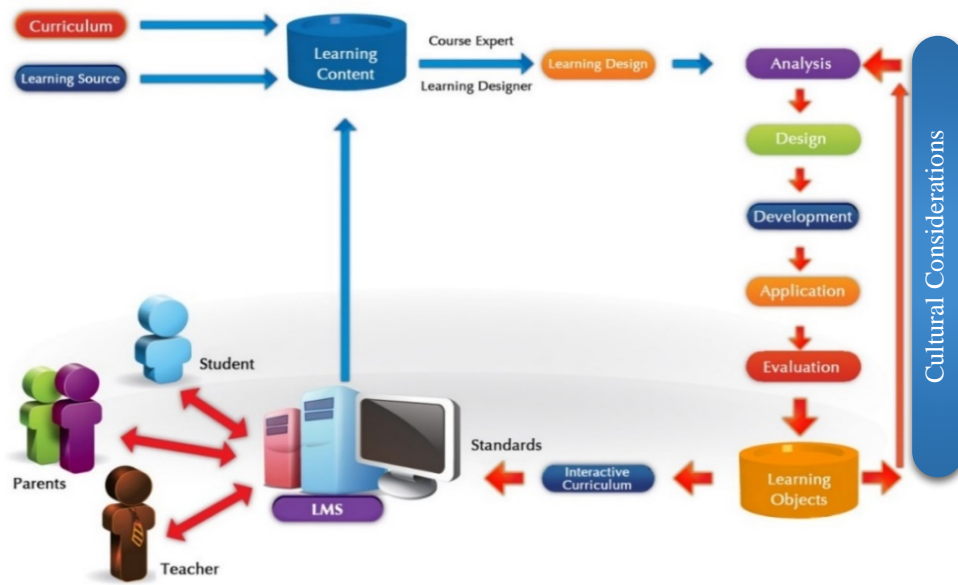


Fig. 3. e-Content Development Model [17].

VI. CONCLUSION AND FUTURE DIRECTION

This research proposes an integrated framework for e-learning implementation that can support a complete or hybrid transformation from traditional teaching methods to online and remote learning. Analysis procedures were chosen to assess the present lockdown situation and the pitfalls and successes that other systems have trialed and recorded. This investigation has also assessed and examined the remote and e-learning structures embraced worldwide and gathered data via a questionnaire. This helped in drawing all the results together to inform a robust implementation framework that, if utilized, can guide Kuwaiti education providers – as with those of neighboring countries – forward in making the most of integrated e-learning. The proposed framework is designed to establish the groundwork that can result in a subsequent dynamic e-learning system in action and aligned with relevant social and cultural contexts. To be efficacious, developers are advised to base their work around e-Content versatility, which should provide the elasticity to connect to the various aspects of e-learning. All possible models and structures can be executed effectively, allow Kuwaiti education providers to make the most of online learning potential without having to worry about the restrictions of lockdown harming professional methods or student progress.

Vital features that impact e-learning uptake as an answer to face-to-face teaching being banned in Kuwait due to pandemic have been analyzed and explored in this paper. The approach has been quantitative in nature, making the most of a large group of respondents gathered from a range of Kuwaiti education facilities. In judging the results, we can reach the firm conclusion that the majority are in favor of digitized learning potential, largely seeing its impact as positive and being keen to overcome any obstacles. However, this does not mean that familiarity with e-learning materials and related technology cannot be upgraded. This study confirms certain issues that prevent some from wholly embracing e-learning,

with institutional and individual issues alongside technological factors. Of the latter, the most essential advances need to occur within LMS, infrastructure, online assistance, communication and correspondence tools, and network bandwidth.

Furthermore, within institutions themselves, professionals need to address overall educational strategies to be integrated with online methods and ensure that effective supervision and managerial support are present, along with training initiatives if necessary. The pandemic, although frustrating, is also the best time to get this right, offering an ideal window for assessing all aspects of the online learning experience and refining them for future generations. Therefore, policymakers in Kuwait should prioritize encouraging the design, development, and implementation of innovative new platforms so all these vital lessons can be learned sooner rather than later, if not before the pandemic is over [61]. Other lessons must include being ready for crisis and recognizing that we have no idea when the next lockdown situation will arrive.

Looking ahead to future research, it would be useful if a similar approach was taken but incorporating other GCC countries, for example, the UAE, Saudi Arabia, Qatar, or Bahrain. Comparatively speaking, these countries likely face similar social and cultural challenges, meaning that researchers could work together across countries to upgrade their methods. Plus, adjusting students' approaches is among the key aims of effective e-learning programs, so there could be a closer focus on how students behave regarding new technology and teachers and other contributing professionals. Investigations that focus on analyzing behavior habits may inform developers how users will respond to future developments in class environments and personal studies. Furthermore, we would advise on assessing e-learning systems and how users are making the most of them. A continued focus on performance issues means further upgrades and remaining up to date with modern potential.

ACKNOWLEDGMENTS

The current research was funded by the Kuwait Foundation for the Advancement of Sciences under project code: "CORONA PROP 95".

REFERENCES

- [1] S. Al-Sharhan and A. Al-Hunaiyyan, "Towards an effective integrated e-learning system: Implementation, quality assurance and competency models. In Digital information management (icdim)," in Seventh International Conference, Macau, 2012.
- [2] P. Serdyukov, "Innovation in education: what works, what doesn't, and what to do about it?," *Journal of Research in Innovative Teaching & Learning*, Vol. 10 No. 1, pp. 4-33, 2017.
- [3] R. Florida, "Mapping the World's Knowledge Hubs, CityLab," 26 Jan 2017. [Online]. Available: http://www.citylab.com/work/2017/01/mapping-the-worlds-knowledge-hubs/505748/?utm_source=feed. [Accessed 3 Feb 2021].
- [4] OECD, *Innovating Education and Educating for Innovation: The Power of Digital Technologies and Skills*, Paris: OECD Publishing, <http://dx.doi.org/10.1787/9789264265097-en>, 2016.
- [5] IGI Global, "IGI Global," 2021. [Online]. Available: <https://www.igi-global.com/dictionary/in-the-pursuit-of-happiness/11494>. [Accessed 2021].
- [6] C. B. Mpungose, "Emergent transition from face-to-face to online learning in a South African University in the context of the Coronavirus pandemic," *Humanit Soc Sci Commun* 7, 113. <https://doi.org/10.1057/s41599-020-00603-x>, 2020.
- [7] R. e. a. Baticulon, "Barriers to online learning in the time of COVID-19: A national survey of medical students in the Philippines," 2020. [Online]. Available: <https://www.medrxiv.org/content/10.1101/2020.07.16.20155747v2>.
- [8] A. Al-Hunaiyyan, R. Alhajri, A. Alzayed and B. Alraqqas, "Towards an Effective Distance Learning Model: Implementation Framework for Arab Universities," *International Journal of Computer Application*. Volume 6, Issue 5, September-October 2016, 2016.
- [9] M. AlKharang, *Factors that Influence the Adoption of e-Learning An Empirical Study in Kuwait*. Phd. Thesis, London: Brunel University London, 2014.
- [10] Academia, "Discussion Forum: Online learning in Light of COVID-19," 29 March 2020. [Online]. Available: <http://dlvr.it/RSnh0X>.
- [11] S. Al-Sharhan, A. Al-Hunaiyyan and H. Al-Sharrah, "A new efficient blended e-learning model and framework for k12 and higher education: Design and," in 2010 fifth international conference, 2010.
- [12] Al-Anbaa, "Opinions of Faculty Members of Online Learning as a reponse to COVID-19 Crises," 10 April 2020. [Online]. Available: https://alanba.com.kw/961334/?utm_source=whatsapp.
- [13] K. Albasayna, *Factors Influencing the Use of E-Learning in Schools in Crises Areas: Syrian Teachers' Perspectives*, Tallinn University of Technology, Estonia, 2016.
- [14] G. Kituyi and I. Tusubira, "A framework for the integration of e-learning in higher education institutions in developing countries," *International Journal of Education and Development using Information and Communication Technology (IJEDICT)*, 2013, Vol. 9, Issue 2, pp. 19-36, 2013.
- [15] A. Al-Hunaiyyan, R. Alhajri and S. Al-Sharhan, "Perceptions and challenges of mobile learning in Kuwait," *Journal of King Saud University – Computer and Information Sciences* Volume 30, Issue 2, pp. 279-289, 2018.
- [16] H. Beetham and R. Sharpe, *Rethinking pedagogy for a digital age: Designing for 21st century learning*, 2 ed., London: routledge, 2013.
- [17] S. Al-Sharhan, T. Al-Sedrawi and H. Al-Sharrah, *E-learning Strategies, and implementation Projects*, Ministry of Education, 1 ed., Kuwait: Ministry of Education, 2010.
- [18] H. Rodrigues, F. Almeida, V. Figueiredo and S. Lopes, "(2019) Tracking e-learning through published papers: a systematic review.," *Comput Education* 136, p. 87–98, 2019.
- [19] V. Arkorful and N. Abaidoo, "The role of e-learning, advantages and disadvantages of its adoption in higher education," *Int J Instruct Technol Distance Learn* 12(1), p. 29–42, 2015.
- [20] W. Horton, "E-learning by design," John Wiley & Sons, 2011.
- [21] J. Zhang, D. Burgos and S. Dawson, "Advancing open, flexible and distance learning through learning analytics.," *Distance Education*, 40:3, DOI: 10.1080/01587919.2019.1656151, pp. 303-308, 2019.
- [22] C. Nerantzi, "The Use of Peer Instruction and Flipped Learning to Support Flexible Blended Learning During and After the COVID-19 Pandemic.," *International Journal of Management and Applied Research*. 7. 10.18646/2056.72.20-013, pp. 184-195, 2020.
- [23] S. Al-Sharhan, A. Al-Hunaiyyan and W. Gueaieb, "Success factors for an efficient blended learning. In Proceedings of the 10th IASTED International Conference on Internet And Multimedia Systems And Applications, pages 77–82," 2006.
- [24] A. Ozadowicz, "Modified Blended Learning in Engineering Higher Education during the COVID-19 Lockdown—Building Automation Courses Case Study," *Education Sciences* 10(10):292. DOI: <https://doi.org/10.3390/educsci10100292>, 2020.
- [25] M. Shahabadia and M. Uplaneb, "Synchronous and asynchronous e-learning styles and academic.," *Procedia - Social and Behavioral Sciences* 176, p. 129–138, 2015.
- [26] A. Nikoubakht and A. Kiamanesh, "(2019) The comparison of the effectiveness of computer-based education and traditional education on the numerical memory in students with mathematics disorder," *J Psychol Sci* 18(73), p. 55–65, 2019.
- [27] C. Liu and F. Long, "The discussion of traditional teaching and multimedia teaching approach in college English teaching," in *International Conference on Management, Education and Social Science (ICMESS)*, 2014.
- [28] A. Bates, *Teaching in a digital age: guidelines for designing teaching and learning for a digital age*, London: Tony Bates Associates Ltd, 2018.
- [29] T. Anderson, "Theories for learning with emerging technologies," *Emerging technologies in distance education* 7(1), p. 7–23, 2016.
- [30] N. Al-Huwail, S. Al-Sharhan and A. Al-Hunaiyyan, "Learning Design for a Successful Blended E-learning Environment: Cultural Dimensions. *Journal of Computer Science*," *INFOCOMP Volume 6 – No. 4*, pp. 60-69, 2007.
- [31] K. Nachimuthu, "Need of e-Content Development on Education," *Education Today, An International Journal of Education & Humanities*, ISSN: 2229-5755, Vol. 03., No.02, pp. 72-80, 2012.
- [32] B. Czerkawski and E. Lyman, "An Instructional Design Framework for Fostering Student Engagement in Online Learning Environments," *TechTrends* 60, <https://doi.org/10.1007/s11528-016-0110-z>, p. 532–539, 2016.
- [33] B. Lockee, "Shifting digital, shifting context:(re) considering teacher professional development for online and blended learning in the COVID-19 era," *Educational Technology Research and Development*, 1-4., 2020.
- [34] S. Al-Sharhan, "Smart classrooms in the context of technology-enhanced learning (TEL) environment: A holistic Approach," in *Transforming Education in the Gulf Region – Emerging Learning Technologies and Innovative Pedagogy for the 21st Century*, London, Taylor & Francis, 2016.
- [35] D. Newton and A. Ellis, "A model for e-learning integration," Honolulu, Hawaii, 2006.
- [36] A. Al-Hunaiyyan, A. Al-Sharhan and R. Alhajri, "A New Mobile Learning Model in the Context of the Smart Classrooms Environment: A Holistic Approach," *International Journal of Interactive Mobile Technologies (IJIM)*. Vol.11_No.3(2017), pp. 39-56, 2017.
- [37] Y. Hsu and Y. Ching, "A Review of Models and Frameworks for Designing Mobile Learning Experiences and Environments," *The Canadian Journal of Learning & Technology*. published by Canadian Center of Science and Education, vol. 41, no. 3, pp. 1-22, 2015.
- [38] B. Cheawjindakarn, P. Suwannathachote and A. Theearongchaisri, "Critical Success Factors for Online Distance Learning in Higher Education: A Review of the Literature," in *Creative Education* 2012, 2012.

- [39] D. O'Doherty, M. Dromey, J. Lougheed and et al. , "Barriers and solutions to online learning in medical education – an integrative review," *BMC Medical Education*18, 130 (2018). <https://doi.org/10.1186/s12909-018-1240-0>, 2018.
- [40] R. Alhajri and A. Al-Hunaiyyan, "Integrating Learning Style in the Design of Educational Interfaces," *ACSIIJ Advances in Computer Science: an International Journal*, Vol. 5, Issue 1, No.19 , January 2016. ISSN : 2322-5157, 2016.
- [41] B. Adeoye, "The Era of Digital Technology in Teaching and Learning in Nigeria Educational Institutions. 2020," in *The Roles of Technology and Globalization in Educational Transformation*, Hershey, PA: IGI Global, 2020, 2020.
- [42] R. Alhajri, S. Al-Sharhan, A. Al-Hunaiyyan and T. Allothman, "Design of educational multimedia interfaces: individual differences of learners," in *Proceedings of the Second Kuwait Conference on e-Services and e-Systems*, Kuwait, 2011.
- [43] V. Varvel, "Master online teacher competencies," *Online Journal of Distance Learning Administration*, Spring 2007. Vol.10., No.1, 2007.
- [44] S. Dasari, "Using Instructional Design in Developing Instructional New Media Materials," in *Proceedings of national seminar on e-learning and its technologies, ELETECH INDIA 2001*, Session 4-4, Paper 16, 2001.
- [45] J. Lee and S. Jang, "A methodological framework for instructional design model development: Critical dimensions and synthesized procedures," *Education Technology Research and Development*,62(6), doi:10.1007/s11423-014-9352-7, p. 743–765, 2014.
- [46] A. Elias and R. Cavana, "Stakeholder Analysis for Systems Thinking and Modelling," 2011. [Online]. Available: <https://www.researchgate.net/publication/253711729>.
- [47] S. Raza, A. Siddiqui and C. Standing, "Exploring Systemic Problems in IS Adoption Using Critical Systems Heuristics," *Systemic Practice and Action Research*, 32. <https://doi.org/10.1007/s11213-018-9467-6>, p. 125–153, 2018.
- [48] J. Demuyakor, "Coronavirus (COVID-19) and Online Learning in Higher Institutions of Education: A Survey of the Perceptions of Ghanaian International Students in China," *Online Journal of Communication and Media Technologies*, 10(3), 2020.
- [49] A. Chi, *Development of the readiness to teach online scale*, University of Denver, 2015.
- [50] M. Girik Allo, "Is the online learning good in the midst of Covid-19 Pandemic? The case of EFL learners. 10," *Jurnal Sinestesia*, Vol. 10, No. 1, April 2020, pp. 1-10, 2020.
- [51] M. Acosta, A. Sisley, J. Ross, I. Brailsford and A. Bhargav, "Student acceptance of e-learning methods in the laboratory class in Optometry," *PLOS ONE* 13 (12), 2018.
- [52] Anbaa, "Benefits of Distance Learning," 1 April 2020. [Online]. Available: https://alanba.com.kw/959659/?utm_source=whatsapp.
- [53] Alanbaa, "Benefits of Distance Learning," 3 May 2020. [Online]. Available: https://alanba.com.kw/959659/?utm_source=whatsapp. [Accessed 3 May 2020].
- [54] M. Bączek, M. Zagańczyk-Bączek, M. Szpringer, A. Jaroszyński and B. Woźakowska-Kapłon, "Students' perception of online learning during the COVID-19 pandemic: a survey study of Polish medical students," 2020. [Online]. Available: <https://www.researchsquare.com/article/rs-41178/v1>.
- [55] J. Biggs, "The reflective institution: Assuring and enhancing the quality of teaching and learning," *Higher Education*, 41(3), p. 221–238, 2001.
- [56] A. Al-Hunaiyyan, S. Al-Sharhan and H. Al-Sharrah, "A New Instructional Competency Model: Towards an Effective E-Learning System and Environment," *International Journal of Information Technology & Computer Science (IJITCS)*, vol. 5, pp. 94-103, 2012.
- [57] S. Al-Sharhan, A. Al-Hunaiyyan, R. Alhajri and N. Al-Huwail, "Utilization of Learning Management System (LMS) Among Instructors and Students," in *Advances in Electronics Engineering. Lecture Notes in Electrical Engineering*, vol 619, Singapore , Springer, 2020, pp. 15-23.
- [58] A. Al-Hunaiyyan, S. Al-Sharhan and R. Al-Hajri, "Prospects and Challenges of Learning Management Systems in Higher Education," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 11, No. 12, <http://dx.doi.org/10.14569/IJACSA.2020.0111209>, pp. 73-79, 2020.
- [59] K. Shelton and G. Saltsman, "Using the Addie model for teaching online," *International Journal of Information And Communication Technology Education*, 2(3), pp. 14-26, 2006.
- [60] U. Maldonado, G. Khan, J. Moon and J. Rho, "E-learning motivation and educational portal acceptance in developing countries," *Online Information Review*. 35, pp. 66-85, 2011.
- [61] A. Al-Hunaiyyan and S. Al-Sharhan, "The Design of Multimedia e-learning Systems: Cultural Considerations.," in *Proceeding of the 3rd International Conference on Singals, Circuits and Systems*, November 6-8, 2009, Djerba, Tunisia, 2009.

Development and Print of Clothing through Digitalized Designs of Natural Patterns with Flexible Filaments in 3D Printers

Jean Roger Farfán Gavancho¹, Wilber Antonio Figueroa Quispe², Dayvis Victor Farfán Gavancho³
Beto Puma Huamán⁴, Victor Manuel Lima Condori⁵, George Jhonatan Cahuana Alca⁶
Universidad Nacional de Juliaca^{1,2,4,5,6}
Instituto Superior Tecnológico Manuel Núñez Butrón³

Abstract—This study proposes clothing development by digitalizing natural patterns with flexible filaments in 3D printers. The motivation to carry out this research was the similarity and evenness features of fractals. For this purpose, three natural exemplars have been selected and subsequently digitized: Snowflake, Honeycomb, and Flower of Life, with line variants and infill density at 12%. The garment was printed with Thermoplastic polyurethane (TPU) and Thermoplastic Elastomer (TPE) in two 3D printers: Anet A8 and M3D Crane Quad. Lastly, the combination of filaments, printers, line variant, and infill density resulted in forty-eight (48) samples. Two tests were carried out on the printed patterns for the research: The elongation and tensile strength test. The elongation test consists of applying a variable force to each exemplar in order to obtain the percentage of its elastic limit before reaching its fracture point. The tensile test applies a variable vertical power to each design to determine how it will behave under particular pressure. Results show that the snowflake pattern with line variant obtained the best performance in the elongation test compared to the tensile test. Subsequently, four clothing samples were printed with TPU and TPE materials on the two printers mentioned above. The garments are composed of twenty-nine (29) pieces respectively which were connected with a 3D pen. Finally, the item of clothing was worn by five volunteers of different sizes, as shown in the following pages.

Keywords—Natural pattern; fractal; garment; digital design; flexible filament; 3D printing

I. INTRODUCTION

Technology advances fast, impacting directly on different industry sectors, such as the Textile Industry. Stereo lithography (S.L.) is an example of this, and it refers to converting 3D data to physical products using 3D printing [1]. 3D modeling is the process of digitalizing an image that the user previously shaped.

On the other hand, with the rise of digital manufacturing, new terms appear; for example, Computer Assisted Manufacturing (CAM), Computer-aided Design (CAD), and Computer-aided Engineering (CAE). Thus there is the convergence of control machines and computer science, transforming fields such as aerospace, naval, and textile industry [2].

It is important to use the fractal theory for the designs of natural patterns that have order and complex geometry of nature since it allows to create and homogenize new base patterns with line and subsequently elaboration of garments. This emphasizes the factors of creativity, the dynamism of the geometric pattern, similarity, and repetition, which are characteristics of human fractals [3]. Furthermore, it helps to the innovation of digitalized designs of natural patterns for their application in garments in the textile industry.

Elements such as the design and basic pattern, digital manufacturing, and 3-D printing make new production processes possible, for example, creating garments and footwear. Danit Peleg is a pioneer in making garments from flexible filament materials in 3D printers[4].

The development of garments from natural snowflake patterns printed in 3D means a new production process, which also creates the challenge of product customization. Besides, the growth of concepts such as Industry 4.0 will allow the user to place an order and print it in a FabLab center [5].

The National University of Juliaca (UNAJ) is an academic institution that promotes research. The Faculty of Textile and Apparel Engineering of UNAJ generates new knowledge through study lines such as "Production processes, design, safety and quality in the Textile and Apparel Industry," [6], which aims to experiment and explore the process of digital manufacturing and 3D printing. Some factors that arise from this research line are cost, time, quality, comfort, resistance, and adaptability.

On the other hand, the investigation leads to generating new lines of research in Textile Engineering, following the improvements of digitized technology, Industry 4.0, and product customization related to garments and footwear. For example, developing new ideas with 3D scanners and software of footwear with padding, analyzing resistance, flexibility, elongation, and doubles in the textile industry. In the same way, analyzing the acceptance and appreciation of the garment and footwear with filaments in 3D printers in potential users at 38000 meters above sea level. There is also the opportunity of examining the adhesion of filaments to products made of alpaca fiber and sheep wool as an added value to innovative fashion in the Textile Industry.

II. RELATED WORKS

For his undergraduate dissertation, [7] printed a garment with a home 3D printer, demonstrating that this technology could be accessible to anyone interested in developing 3D printing. Peleg [6] was inspired by the work of Eugene Delacroix in "Liberty Leading the People," for which she adapted triangular geometric figures as a primary natural pattern and subsequently printed with biopolymer polylactic acid (PLA) and soft PLA.

In [8], several pieces of clothing were printed using Filaflex filament (a variant of TPE), which is differentiated by its flexible nature. Her work consisted of five garments that stand out the mesostructured cellular pattern proposed by Andreas Bastian, characterized by the hexagonal geometric figure.

In [1], [9] produced a 3D printed garment using the natural pattern of Brain coral, which is defined by a hexagonal geometric figure. The design was digitalized with Rhinoceros and Grasshopper software to create the surface of such clothes. Additionally, the author used PLA, ABS (Acrylonitrile Butadiene Styrene), and Filaflex filaments to perform bending tensile tests and determine the most suitable material for developing the garment.

In [10], Research on garment pattern design based on fractal graphics, scholars generated graphics using two types of fractals: floral graphics and artistic geometric graphics, subsequently applied to the garment. The leading software was MATLAB, supported by Adobe Photoshop. The fractal patterns were applied on silk scarves with digital printing technology. It generates complex and new designs through creativity

In [11], A study on fractal patterns for the textile design of the fashion design; drawings were made based on Julia sets pattern and algorithms. This new methodology generates designs for the Textile Industry. The software used to create the algorithms was JAVA, allowing to generate new and complex designs that traditionally could not be drawn.

In [12], authors made tensile and elongation tests on six PLA-based filaments with some additive materials for 3D printing. The tests printed specimens with 100% filling and with dimensions of 115 mm x 20 mm and 4 mm thickness were included. The printer used was Cetus MKII and an Instron 5566 Universal Testing Machine. Specimen specifications are set up according to ISO 3167 1994. The filament that obtained the best performance in the tensile test was PLA additive with metal obtaining 121.36 N. and the elongation test with PLA technical additive filament with 20.16%.

In [13], Determination of elastic properties of polymeric pieces constructed by 3D printing, subjected to bending; authors analyzed the elastic properties (elongation) of PLA, Nylon, and HIPS filaments. Printed specimens were developed with 80 mm x 10 mm dimensions, a thickness of 4mm, and a percentage of filling at 100%. The 3D printer model used was a Prusa Mendel M90. The novelty of this test is the angle of the infill line and the specimen wall with four combinations: 0°/0° (long line infill), 45°/-45° (rhombic infill), 0°/90° (squared infill), as well as the change of specimen position for vertical and horizontal printing. The 0°/0° angle specimens obtained a

better result with an average of more than 3% elongation. The 0°/90° angle vertically printed specimens performed better with an average of more than 3% almost as well as the 0°/ 0° angle horizontally printed specimens.

In [14], the influence of infill parameter on the mechanical resistance in 3D printing, using the Fused Deposition Modeling method; the research objective was to determine the influence of infill on the tensile strength, tested under ABS filament. The printer used was a Makerbot Replicator 2X, and printed specimens with different filler percentages, starting from 0% to 5%, were selected for this test. The standard test method applied was the ATSM D638-10 and the material tester utilized for the tensile test was a Gunt Hamburg WP 310 of 50 kN. The results obtained were: 34.57 MPa with a percentage of filler at 100% and 14.62 Mpa with a percentage of filler at 0%. Therefore, the conclusion establishes that the higher the percentage of filler, the higher the tensile strength. However, for the manufacture of garments, the traction property must go along with the elongation feature.

In [15], Tensile strength of commercial polymer materials for fused filament fabrication 3D printing; the experiment was performed with three flexible filaments: NinjaFlex, Semiflex, and Nylon. Besides, standard filaments such as ABS, HIPS, Polycarbonate, and T-Glass were also utilized. The printed specimens had a 60 mm x 13 mm dimension and a thickness of 3.2 mm using the ASTM D638 standard. The material tester was an INSTRON 4206 with a capacity of 10kN for rigid filaments and INSTRON 4210 for flexible filaments due to their higher elasticity compared with standard filaments. The conclusion states that standard filaments perform better in the tensile test than flexible filaments. Polycarbonate reached the highest value with 2041.64 N, while Nylon reached a value of 1102.07 N. Lastly, NinjaFlex flexible filament was the lowest performer, as shown in its result: 161.88 N.

III. THEORETICAL BACKGROUND

A. Fractals

In 1975, Benoit Mandelbrot first introduced the word "fractal," derived from the Latin term "fractus" [16]. Two meanings are attributed to this term; the first is to break into pieces, and the second means irregular. However, Mandelbrot gave another definition to fractals, describing them as geometric objects with characteristics of internal and invariable similarity structure. Besides, a fractal repeats its basic structure at different scales to increase it or reduce it, being appreciated in various forms of nature such as trees. Two characteristics are attributed to fractals:

- They are too irregular; therefore, they cannot be described with traditional geometric patterns.
- They are self-similar; in other words, they are small copies of the exact figure.

In [17], defines the fractal figure as "Irregular and complex set of structures formed through computational and mathematical algorithms, with basic figures such as points, straight lines and others of traditional mathematics," while [18] defines it as "very irregular shape, broken or fragmented at any scale, but with a distinctive feature or pattern."

It is essential to differentiate between two concepts: "fractal set" and "natural fractal." The former is described as "a set of characteristics that can be defined through mathematics" while the latter is defined as a "natural object with regularity; for example clouds, trees or others" [17].

B. Natural Patterns

Nature has copious fractal patterns. [19] describes that there is no chaos in nature, and on the contrary, there are levels of order with their complex structures, which in nonlinear mathematical terms it is defined as self-ordering. These natural patterns can be found in different locations, such as snowflakes, honeycombs, ground cracks, peacock feathers, lightning, clouds, et. al.

In [20], the author explains that fractals must be understood in different ways, being one of them defined by Euclidean geometry. Said book [20] provides the reader with several examples such as galaxies, swirls, snowflakes, islands, branches, trees, lakes, cosmic dust, the coast of Great Britain, lunar craters, and others.

1) *The fractal geometry of nature*: It is complex to describe nature within the natural fractal geometry since its irregular and fragmented character could reach an endless level of complexity [21].

Euclidean Geometry discards the study of nature due to its amorphous essence. However, in addition to his fractal theory, Mandelbrot developed the Fractal Geometry of Nature, which permits the description of the irregular and fragmented patterns of nature and defines two relevant terms: "randomness" and "statistics" [20].

2) *Human fractals*: The concept refers to visual expressions created by people [19].

Human fractals are characterized by creativity, color combination, mystery, spirit, dynamism, and the mixture of these elements are said fractals [20]. These human creations can occur in diverse art fields such as painting, sculpture, music, film, digital art, and others. Among the best-known examples is Leonardo Da Vinci's work: "Drawing of a flood," where irregular and complex shapes of clouds and water are observed, as well as whirlpools of different sizes. Moreover, "The Big Wave" by Katsushika Hokusai [20] stands out as one of the first fractal paintings which focuses on elements such as the sea, waves, clouds, and treetops.

3) *Koch snowflake*: It has a complex aspect considered as natural fractal geometry [20]. It is usually compared to Great Britain's coasts due to its cascading shape, although the latter possesses a rudimentary pattern. The Koch Snowflake could be considered a fractal of simple nature; however, Euclidean Geometry defines it as complex.

4) *Fractal and the algorithms of nature*: Euclidean Geometry calculates approximations of nature's fractal geometry, helped by Computer-Aided Engineering, denominating this process as "Fractal Geometry" [22]. Fractal Geometry and Euclidean Geometry differ when working with

circumferences and infinite repetitive processes, producing one of two situations:

- When increasing, the item can become a straight line or.
- When reducing, it will be lost at some point.

For Euclidean Geometry, the item can lose the original form, but for Fractal Geometry, it has the term of self-similarity, being noticed in reduction or enlargement operations. In order to get a diagram, it is necessary to use algorithms.

5) *Biophilic design*: Biophilia is defined as "the relationship between human beings and nature" [23]. The biophilic design incorporates patterns to the architecture (houses, buildings, interiors, et al.), which can be habitable by the human being as living space. This design presents more harmonic and non-traditionalist techniques and helps people in different aspects such as health through stress reduction, as shown by research conducted in [23]. Among other benefits, the article mentions the perception through cognitive function and development of creativity and intellect. A notable example of biophilic design is The Sphinx in Egypt. Nowadays, many biophilic designs can be found incorporating living organisms such as plants and animals or other replication designs from nature.

6) *Fractals in textile industry*: Computer-Aided Design (CAD) has helped in different areas to develop new production processes, being one of them the Textile Industry. With unique designs generated from creativity and productivity, CAD realizes automation, time reduction, customization, and product valuation. Thus, the use of fractals in the Textile Industry [24] has served to associate the idea of fractal as a product in the creation of garments, fabrication, printing, and others. Besides, with the implementation of algorithms, fractals generate products such as carpets, fabrics, leather, et. al.

The use of technology in the Textile Industry is not new. In [25], some trends were generated aided by software and simulators showing that technology can create complete collections and even individual and customized designs.

Observing the use of fractals in computer science is fascinating. However, it can be slightly complex because the fractal must be pre-designed, applying the image, processing techniques, algorithms, and others. For the design, it is possible to use digitalization or scanning techniques, color theory, deleting damaged areas with different software. It is also possible to select entirely or partially the work area and modify the size by reduction or enlargement.

Although there are programs that support the design, generate the image, program the algorithm, or simulators such as Corel Draw, they must be compatible. This is still a disadvantage since they operate independently, producing the risk of image loss which can occur when importing from one program to another, and this in the textile field could mean a significant loss.

C. 3D Printing

In [9], 3D printing is defined as "the sequential process to create objects in three dimensions from digitalized data, through the collection of material on a platform." On the other hand, [26] indicates that it is "a process where materials are combined to create objects in physical form from a digital design."

Domestic and industrial 3D printers have different uses depending on their application. Some processes are new and help to improve experimental areas, such as building housing, producing human tissues, and manufacturing spare parts in highly isolated places [27].

The process of layers resin salification was done through ultraviolet light. By the year 2000, 3D printing adds concepts such as additive manufacturing of casting patterns (Rapid Casting), production tools through injection molding (Rapid Tooling), and the gathering of production parts (Rapid Manufacturing). Due to these improvements, nowadays, 3D Printing is accessible and easier to use [26], emphasizing the digitalization of objects [28].

1) *Printing techniques*: According to the definitions reviewed, there are four techniques in 3D printing:

a) *Additive*: It is the union of various materials, techniques, equipment, et al.

b) *Subtractive*: Refers to the exclusion of material from the original design. These can be solid objects.

c) *Transformation*: Change of the material by other techniques and methods.

d) *Hybrid*: Combination of materials, processes, procedures, equipment, et al.

2) *Rapid prototyping*: It is a process to create new products, attempting to know the market's impact, and collecting answers for manufacturing. Rapid prototyping is an automated way to elaborate layers of different materials. It is commonly defined as "Layer Manufacturing, that allows obtaining prototypes with a wide range of materials, regardless the shape and geometric complexity, in a short time and without using other tools" [29].

D. Digital Manufacturing

In [27], it is defined as "the process of developing objects from digital files, using an equipment (machine) controlled by a computer."

Digital Manufacturing is a broad and technical process focused on the social aspect since many advantages can be obtained. Some of them customize products, reduce production costs, manufacture complex objects and nanometric precision provided by the software and programming languages that support it.

Nowadays, digital manufacturing is involved in different areas of knowledge directly or indirectly. For this purpose, it needs the complement of computer-controlled machines such as a Computer Numerical Control (CNC), which impacts different industries such as automotive, textile, or in some cases in people. It is open-source, promoting the exchange of

information between development groups. However, it is possible that digital manufacturing becomes lucrative in the future since it will be shaped according to the needs of organizations and individuals [27].

IV. METHODOLOGY

A. Research Sample

This research experiments with three proposed natural patterns defined by a hexagonal geometric design having the line variant and infill at 12%. Two types of 3D printers have been utilized with two different flexible filaments described in the following (Table I).

B. Equipment

The equipment and tools used to conduct this research are described below.

1) *3D printers*: The 3D printers employed for this study are Anet A8 and M3D Crane Quad (see Fig. 1 and 2). Their characteristics are described in Table II.

TABLE I. RESEARCH SAMPLES. DISTRIBUTION OF THE 48 SAMPLES BY FILAMENT, PATTERN, AND INFILL VARIANT

Filament	Natural Pattern	Variant	3D Printer		Total
			A8 Anet	M3D Crane Quad	
TPU	Snowflake	Line	2	2	4
		Infill	2	2	4
	Flower of life	Line	2	2	4
		Infill	2	2	4
	Honeycomb	Line	2	2	4
		Infill	2	2	4
TPE	Snowflake	Line	2	2	4
		Infill	2	2	4
	Flower of life	Line	2	2	4
		Infill	2	2	4
	Honeycomb	Line	2	2	4
		Infill	2	2	4

TABLE II. 3D PRINTERS FEATURES

Feature	3D Printer	
	Anet A8	M3D Crane Quad
Nozzle Diameter	0.4 mm.	0.35 mm.
Build Volume	220 x 220 x 240 mm.	200 x 200 x 230 mm.
Printing Speed	40 -120 mm / s	hasta 80 mm / s
Filament diameter	1.75 mm.	1.75 mm.

2) Dynamometer (see Fig. 3)

- Model: Thread Dynamometer model 848 from Instruments J. Bot S.A.
- Maximum Capacity: 10 kg.
- Elongation: 0.1% over 500mm.
- Upper clamp for wire
- Lower clamp for driving roller

3) Intelligent 3D Printing Pen (see Fig. 4)

- Brand: SUNLU 3rd generation.
- Model: SL-300.
- Temperature: ABS 180~210°C and PLA 160~180°C.
- Filament diameter: 1.75mm.



Fig. 4. Smart 3D Printer Pen SUNLU used to Join the 29 Pieces of the Garment Printed with the Snowflake Pattern.

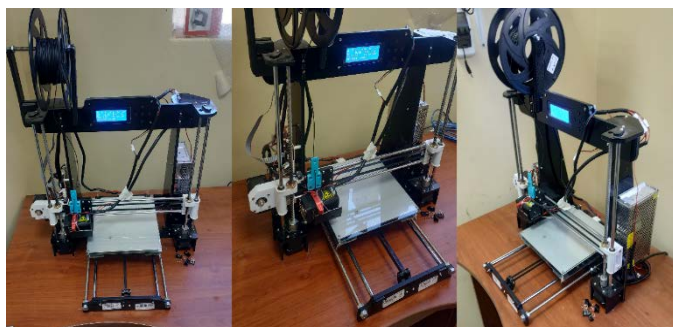


Fig. 1. Anet A8 3D Printer with 0.4 mm. Nozzle Diameter and Modified to Print Flexible Filaments.



Fig. 2. M3D Crane Quad 3D Printer with 0.35 mm. Nozzle Diameter and the Capacity to Combine Four Colors of Filaments.



Fig. 3. Dynamometer Model 848 from J. Bot S.A. used in Tensile and Elongation Tests of the Printed Natural Patterns.

C. Software

The software programs utilized for this research are described in the following.

1) *Rhinoceros 6*: 3D design and modeling software used to shape the patterns of the garment.

2) *CLO 3D*: Software for fashion design and simulation in 2D and 3D, which allowed the validation of dimensions and simulation of the garment with the selected pattern.

3) *Voxelizer 2.1*: A 3D printing software that permitted the generation of the GCODE file.

4) *Plug-in*: The author downloaded three plug-ins and used them in the Rhinoceros design and simulation. These are Trace, SymmetryArrow, and Trim Arrow.

D. Procedures Research

1) *Natural Patterns*: The design of the main pattern takes natural forms as a reference. Such models are defined by hexagonal geometry, permitting an easy union through the network.

While in [4], the author was inspired by the mesostructured cellular pattern by Andreas Bastian; in [9], the main design was the Brain coral adjusted to hexagonal geometry (see Fig. 5).

This research used three natural hexagonal patterns: "Snowflake," "Honeycomb," and "Flower of Life" (see Fig. 6). The dimensions of the natural patterns are described in the Table III. The mentioned patterns have been modified to continue with their digitalization. Lastly, corrections were applied to the line variant and infill density (see Fig. 7, 8 and 9).

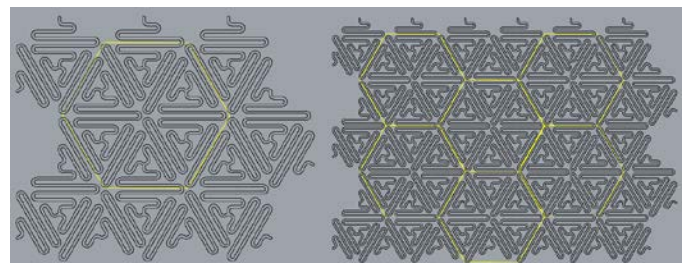


Fig. 5. Mesostructured Cellular Pattern with Hexagonal Geometry by Andreas Bastian, Showing the Union of the Patterns.



Fig. 6. Natural Patterns "Snowflake," "Honeycomb," and "Flower of Life" used for the Traction and Elongation Tests and Subsequent Garment Elaboration.

TABLE III. MEASUREMENTS OF PROPOSED NATURAL PATTERNS

Natural Pattern	Variant	
	Line	Infill
Snowflake	13.2 x 10.8 x 0.15 cm.	13 x 9.4 x 0.15 cm.
Honeycomb	10.9 x 7.5 x 0.15 cm.	10.9 x 8.9 x 0.15 cm.
Flower of life	11.7 x 10.7 x 0.15 cm.	11.3 x 10.3 x 0.15 cm.

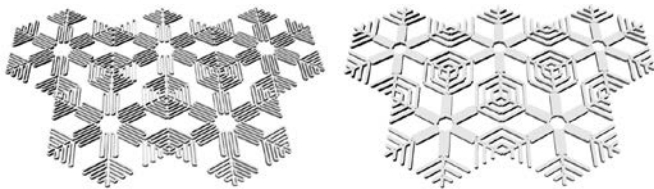


Fig. 7. Snowflake Pattern Line Variant and Infill Density, Digitalized through Rhinoceros 6 Software.

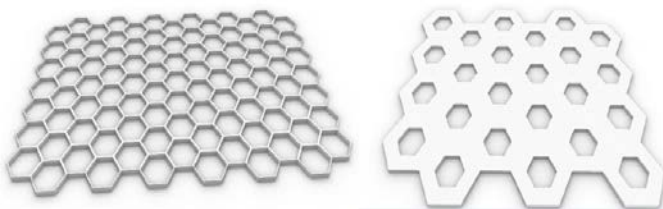


Fig. 8. Snowflake Pattern Line Variant and Infill Density, Digitalized through Rhinoceros 6 Software.

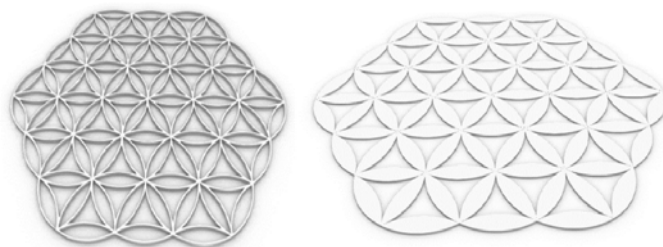


Fig. 9. Flower of Life Pattern Line Variant and Infill Density, Digitalized through Rhinoceros 6 Software.

2) *Printing process:* For this research, the two 3D printer filaments used were TPU and TPE. The combination of filaments, printers, line variant, and infill density resulted in forty-eight (48) samples printed. While the first sample was named "P1", the second unit was called "P2". The technical specifications of the printing are described in the Table IV.

TABLE IV. TECHNICAL SPECIFICATIONS OF THE 3D PRINTER PARAMETERS CONFIGURATION FOR THE CREATION OF THE PATTERNS

Parameter	3D Printer	
	Anet A8	M3D Crane Quad
Extruder temperature	240°C	240°C
Hotbed temperature	70°C	70°C
Quality Print	60%	60%
Infill	12%	12%
Raft	No	No
Retraction	No	No
Support	No	No
Resolution	0.22 mm.	0.22 mm.

3) *Tests:* For the purpose of this research, two tests were carried out on the samples: The elongation and tensile strength test. The elongation test consists of applying a variable force to each exemplar in order to obtain the percentage of its elastic limit before reaching its fracture point. The tensile test applies a variable vertical power to each design to determine how it will behave under particular pressure (see Fig. 10, 11, 12, 13, 14, and 15).



Fig. 10. Tensile and Elongation Tests of Snowflake Pattern with Line Variant, Printed with TPU and TPE Flexible Filaments.



Fig. 11. Tensile and Elongation Tests of Snowflake Pattern with Infill, Printed with TPU and TPE Flexible Filaments.



Fig. 12. Tensile and Elongation Tests of Honeycomb Pattern with Line Variant, Printed with TPU and TPE Flexible Filaments.



Fig. 13. Tensile and Elongation Tests of Honeycomb Pattern with Infill, Printed with TPU and TPE Flexible Filaments.



Fig. 14. Tensile and Elongation Tests of Flower of Life Pattern with Line, Printed with TPU and TPE Flexible Filaments.



Fig. 15. Tensile and Elongations Tests of Flower of Life with Infill, Printed with TPU and TPE Flexible Filaments.

4) *Digitalization of the garment:* The performed tests showed that the most optimal model to use for the research was the snowflake pattern.

Subsequently, using Rhinoceros 6, the selected pattern was placed in twenty-five columns with thirteen and fourteen repetitions. The latter created a rectangular area of 95.65 cm. x 61.23 cm. x 0.15 cm., ideal to proceed with the garment's design (see Fig. 16).

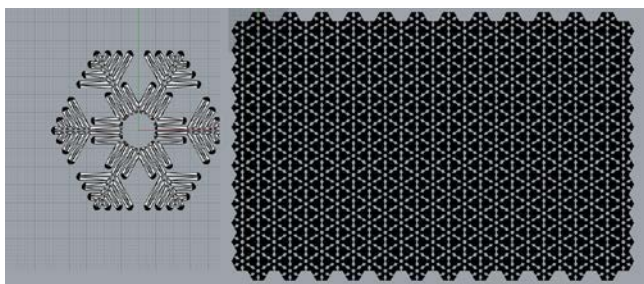


Fig. 16. Snowflake Pattern and Column Sorting, Simulating Fabric Cutting and Subsequent Division into 29 Pieces.

The rectangular area formed by the pattern was imported in image format (PNG) to the CLO 3D Software. Subsequently, the garment's necessary sections were delimited in the mentioned software, as shown in Fig. 17. Size "M" was taken as a reference since it is common among clothes of women (see Fig. 17).

With the model entirely designed, the authors returned to Rhinoceros and deleted the patterns that were not needed (see Fig. 18). The author deleted column 25 to obtain regular patterns to print, as shown in Fig. 18.

The current workspace was exported again in PNG to CLO 3D to perform the garment simulation with TPU and TPE filaments (see Fig. 19).

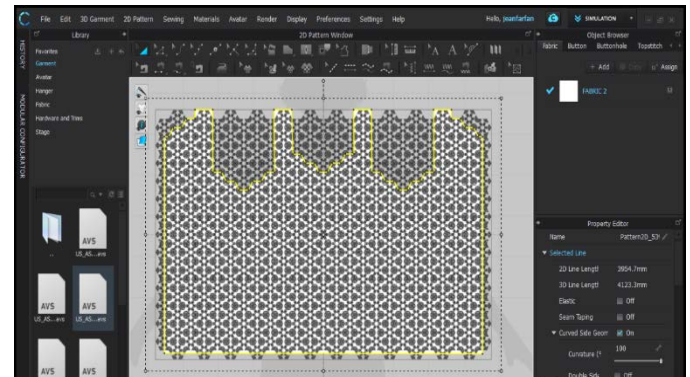


Fig. 17. Design of the Garment on the Rectangular Working Area in CLO 3D Software in Order to Delete Areas of Pattern Located Outside the Garment Area in the Yellow Region.

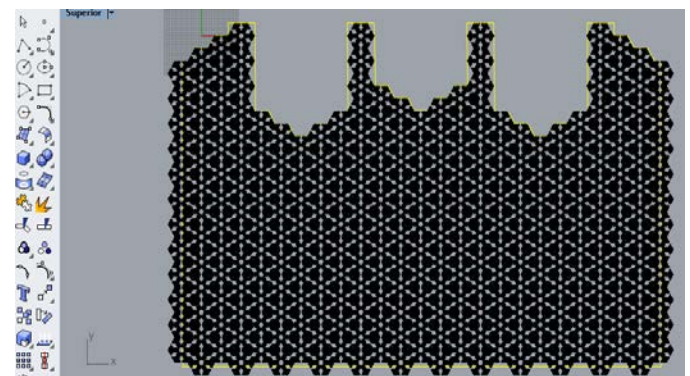


Fig. 18. Deleting Non-Essentials Areas of the Garment in Rhinoceros 6 Software to Simulate the Real Garment.



Fig. 19. Garment Simulation with TPE and TPU Flexible Filaments in CLO 3D Software for a Woman of Size M.

Due to the printers' size of printing area used for this study, the whole pattern was divided into 29 pieces of 20 cm x 20 cm. (see Fig. 20). Such parts were named with a number from left to right and from the bottom to top. The fragments from 1 to 18 are equal, simplifying exporting to GCODE format (see Fig. 21).

Fragments from 1 to 18 were joined initially, continuing with sections from 19 to 29 to obtain a single item. Lastly, each edge of this single piece was joined together (see Fig. 22 and 24), forming the mentioned garment (see Fig. 26 and 27). The printed pieces' union was achieved using the 3D pen applying its corresponding filament (TPE and TPU) inside the garment at 185 C° (see Fig. 23 and 25).

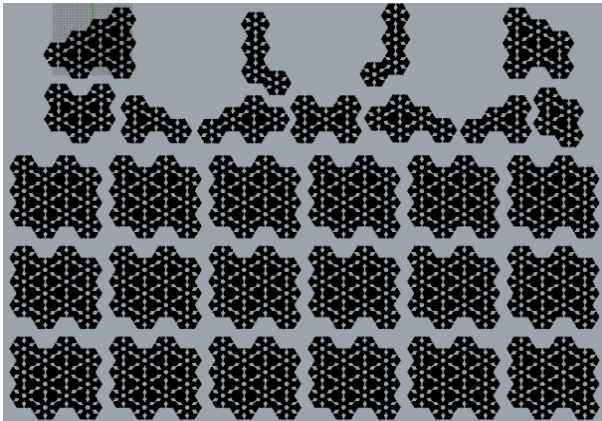


Fig. 20. Preparation of 29 Pieces of Garment for 3D Printing Exported in STL File Format.

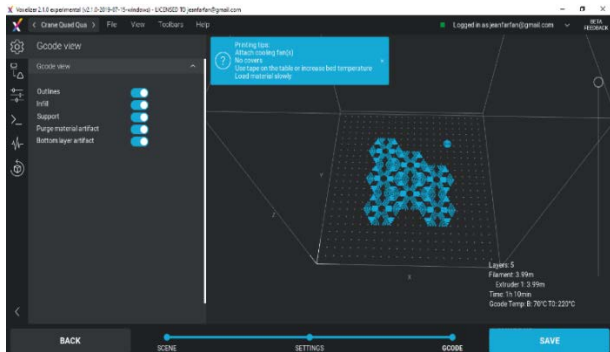


Fig. 21. Generation of the GCODE File through Voxelizer 2.0 Software for 3D Printing.

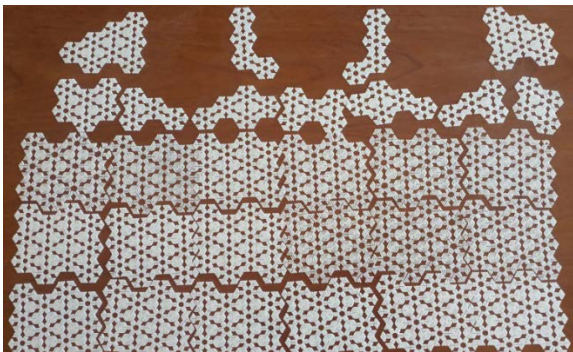


Fig. 22. Printed Garment's Parts with TPU Flexible Filament, Ordered from 1 to 29, from Left to Right, and from Bottom to Top.

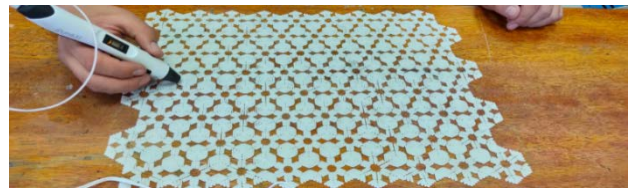


Fig. 23. Union of garment parts with TPU Filament using the 3D Smart-Pen.

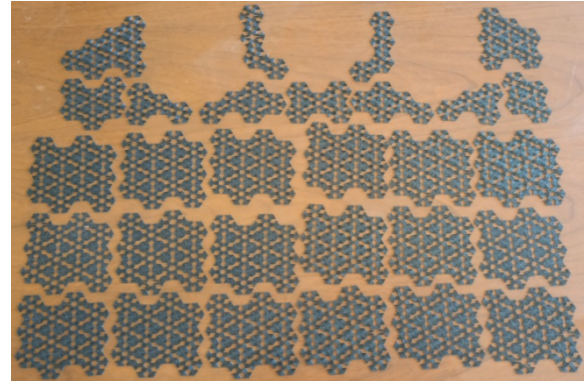


Fig. 24. Printer Garment Pieces with TPE Flexible Filament, Ordered from 1 to 29, from Left to Right, and from Bottom to Top.

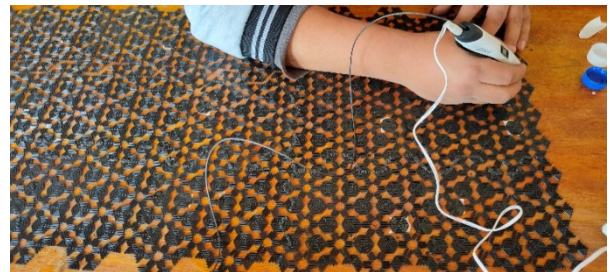


Fig. 25. Union of Garment Parts with TPE Filament using the 3D Smart-Pen.

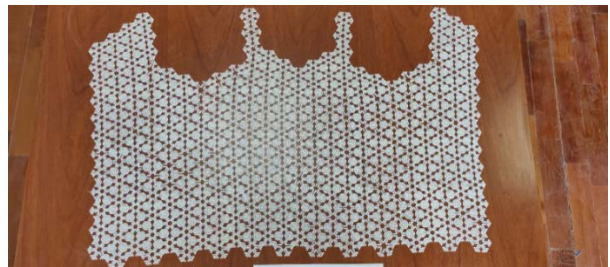


Fig. 26. The Garment used in the Research Printed with TPU Flexible Filament after Completing the Process of Joining Parts with the 3D Smart-Pen.

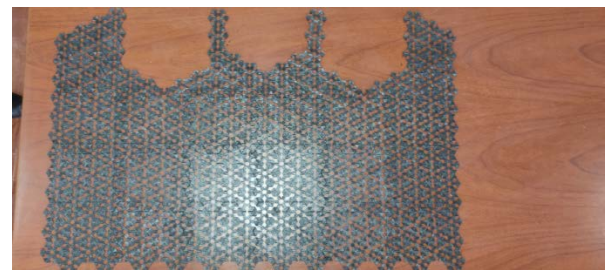


Fig. 27. The Garment used in the research Printed with TPE Flexible Filament after Completing the Process of Joining Parts with the 3D Smart-Pen.

5) *Presentation of the garment*: Volunteers of S, M, L, and XL size, were asked to model the official presentation's garment. The volunteers wore the printed item of clothing, combining it with a skirt, a jacket, and jeans, as shown in Fig. 28, 29, 30 and 31.

Tables V, VI, and VII shows garments time and weight in both 3D printers and flexible filaments.

TABLE V. STATISTICAL DATA OF PRINTED PIECES WITH TPU FLEXIBLE FILAMENT

Piece	Dimensions (in cm.)	3D Printer			
		A8 Anet		M3D Crane Quad	
		Time (h:m)	Weight (g.)	Time (h:m)	Weight (g.)
1	16.37 x 15.35	01:42	15.74	01:49	10.94
2	16.37 x 15.35	01:45	16.17	01:51	11.44
3	16.37 x 15.35	01:44	16.06	01:48	9.45
4	16.37 x 15.35	01:42	15.91	01:52	12.96
5	16.37 x 15.35	01:43	15.99	01:51	11.47
6	16.37 x 15.35	01:44	16.20	01:49	10.63
7	16.37 x 15.35	01:43	15.83	01:50	12.40
8	16.37 x 15.35	01:44	16.01	01:49	11.04
9	16.37 x 15.35	01:43	15.91	01:49	11.45
10	16.37 x 15.35	01:43	16.00	01:49	11.43
11	16.37 x 15.35	01:44	16.34	01:48	10.56
12	16.37 x 15.35	01:45	16.46	01:47	8.95
13	16.37 x 15.35	01:44	16.35	01:47	9.78
14	16.37 x 15.35	01:48	16.22	01:50	12.75
15	16.37 x 15.35	01:44	16.46	01:50	12.06
16	16.37 x 15.35	01:46	16.12	01:51	12.52
17	16.37 x 15.35	01:47	16.00	01:53	13.89
18	16.37 x 15.35	01:47	16.31	01:51	11.29
19	12.60 x 10.98	00:50	7.82	00:58	5.62
20	12.60 x 8.79	00:32	5.04	00:39	4.00
21	16.38 x 8.79	00:41	5.78	00:50	4.76
22	12.60 x 8.79	00:41	6.49	00:49	4.81
23	16.38 x 8.79	00:41	6.36	00:50	4.80
24	12.60 x 8.79	00:32	5.53	00:39	3.34
25	8.24 x 10.92	00:32	4.58	00:39	4.14
26	16.37 x 13.16	01:07	9.30	01:21	7.87
27	8.82 x 15.34	00:32	4.77	00:39	3.92
28	8.82 x 15.34	00:32	5.09	00:39	3.52
29	12.60 x 13.16	00:58	8.62	01:10	6.65

TABLE VI. STATISTICAL DATA OF PRINTED PIECES WITH TPE FLEXIBLE FILAMENT

Piece	Dimensions (in cm.)	3D Printer			
		A8 Anet		M3D Crane Quad	
		Time (h:m)	Weight (g.)	Time (h:m)	Weight (g.)
1	16.37 x 15.35	01:41	16.56	02:02	14.10
2	16.37 x 15.35	01:42	16.87	02:03	14.24
3	16.37 x 15.35	01:40	16.42	02:02	14.19
4	16.37 x 15.35	01:39	14.01	02:01	14.00
5	16.37 x 15.35	01:40	16.45	02:03	14.28
6	16.37 x 15.35	01:39	16.20	02:02	14.25
7	16.37 x 15.35	01:41	16.56	01:59	11.21
8	16.37 x 15.35	01:41	16.59	02:01	13.98
9	16.37 x 15.35	01:40	16.34	02:03	14.34
10	16.37 x 15.35	01:42	16.73	02:00	13.26
11	16.37 x 15.35	01:43	16.62	02:02	14.15
12	16.37 x 15.35	01:40	16.31	02:02	14.14
13	16.37 x 15.35	01:43	16.72	02:04	14.64
14	16.37 x 15.35	01:41	16.40	02:02	14.26
15	16.37 x 15.35	01:41	16.41	02:02	14.16
16	16.37 x 15.35	01:41	16.41	02:02	14.21
17	16.37 x 15.35	01:39	16.23	02:02	14.26
18	16.37 x 15.35	01:40	16.27	02:03	14.57
19	12.60 x 10.98	00:50	8.47	01:01	6.71
20	12.60 x 8.79	00:32	5.57	00:39	4.75
21	16.38 x 8.79	00:41	7.07	00:50	5.96
22	12.60 x 8.79	00:41	7.12	00:49	5.94
23	16.38 x 8.79	00:41	7.01	00:50	5.93
24	12.60 x 8.79	00:32	5.60	00:39	4.76
25	8.24 x 10.92	00:32	5.63	00:39	4.75
26	16.37 x 13.16	01:08	11.10	01:22	9.48
27	8.82 x 15.34	00:32	5.61	00:39	4.76
28	8.82 x 15.34	00:32	5.60	00:39	4.76
29	12.60 x 13.16	00:58	9.80	01:11	8.29

TABLE VII. STATISTICAL DATA OF GARMENTS

Filament	3D Printer			
	A8 Anet		M3D Crane Quad	
	Time (h:m)	Weight (g.)	Time (h:m)	Weight (g.)
TPU	38:56	359.46	42:07	258.44
TPE	37:52	372.68	45:53	318.33

V. RESULTS

A. Elongation Test

In Table VIII, shows the results obtained from the garment design's primary test.

In Table IX, the Elongation test, the authors executed an analysis of variance (ANOVA), which was significant and applicable for filament and pattern factors (see Fig. 32).

TABLE VIII. RESULTS OF ELONGATION TEST

Filament	Natural Pattern	3D Printer			
		A8 Anet		M3D Crane Quad	
		P1	P2	P1	P2
TPU	Snowflake Line	166.90	171.20	197.00	186.60
	Snowflake Infill	22.80	18.50	22.20	16.50
	Flower of Life Line	31.50	36.30	33.00	37.00
	Flower of Life Infill	45.00	21.80	38.20	20.70
	Honeycomb Line	20.90	27.60	16.90	34.60
	Honeycomb Infill	21.70	14.00	19.50	13.20
TPE	Snowflake Line	147.30	160.80	117.40	161.60
	Snowflake Infill	15.30	18.40	17.90	15.20
	Flower of Life Line	28.90	24.60	27.80	26.90
	Flower of Life Infill	26.60	23.00	21.50	23.60
	Honeycomb Line	12.90	24.70	22.10	27.40
	Honeycomb Infill	19.40	9.90	17.10	13.30

TABLE IX. RESULTS OF ANOVA REGARDING ELONGATION TEST

Source	df	Sum of Squares	Mean Square	F	Sig.
Filament (A)	1	1102.083	1102.083	14.254	0.001
Pattern (B)	5	132604.819	26520.964	343.019	0.000
A x B	5	1444.249	288.850	3.736	0.008
Error	36	2783.385	77.3160		
Total	47	137934.537			

Box plot of means regarding elongation test

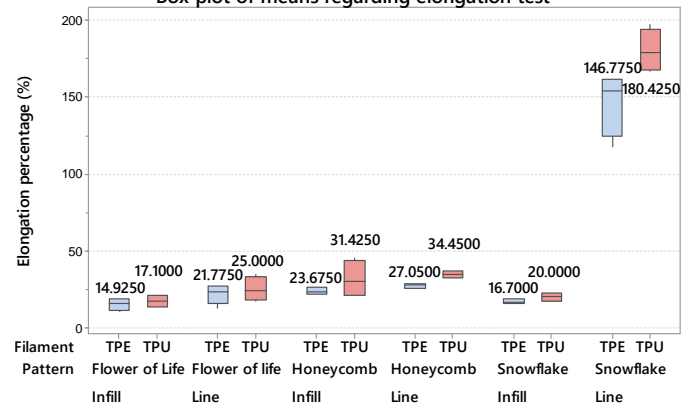


Fig. 32. Box Plot of Elongation Test Means Regarding Pattern and Flexible Filament.



Fig. 28. Volunteers Wearing the Garment Printed with TPU Filament in the Anet A8 Printer.



Fig. 29. Volunteers Wearing the Garment Printed with TPU Filament in the M3D Crane Quad Printer.



Fig. 30. Volunteers Wearing the Garment Printed with TPE Filament in the Anet A8 Printer.



Fig. 31. Volunteers Wearing the Garment Printed with TPE Filament in the M3D Crane Quad Printer.

Flexible filament factor (A). - The F-value = 14.254 leaves on the right a p-value of 0.001, in other words, a significant value. It can be concluded that the design is significant and the inclusion is successful. Thus, the elongation test depends on the filament factor when making the garment.

Pattern factor (B). - The F-value = 343.019 leaves on the right a p-value of 0.000, which is interpreted as a significant value. The result suggests that some pattern significantly influences the elaboration of the garment. In other words, there are significant differences between designs.

A x B. - This row captures the influence of all filament and pattern interactions. The value of F=3.736 leaves on the right a p-value of 0.008, lower than the 5% significance level. There is sufficient evidence that there is an interaction between filament and pattern, or in other words, the studied factors are dependent on each other.

Error. - This row refers to the variance of the dependent variable not explained by the model.

Total. - This row shows the variance observed in the dependent variable caused by all the factors.

However, the ANOVA does not determine which treatment has the most significant average increase. A T-test was performed between flexible filament (A) and pattern(B) to find the mentioned value.

Flexible Filament (A). Table XII show the results of the T-test ($P \geq 0.457$). There is no statistical difference between the TPU and TPE filament during the elongation test (see Table X and Fig. 33).

Pattern (B). Table XI and Fig. 34 show the results of Duncan's new multiple range test ($P \leq 0.05$). It shows exemplars' performance for the Elongation test where the Snowflake pattern (163.6%) has a higher average score, indicating that it is statistically different from other designs.

TABLE X. INDEPENDENT SAMPLES OF ELONGATION REGARDING FLEXIBLE FILAMENT

Filament	N	Mean	Std. Deviation	Std. Error Mean
TPU	24	51,4000	59,73579	12,19352
TPE	24	41,8167	48,79419	9,96007

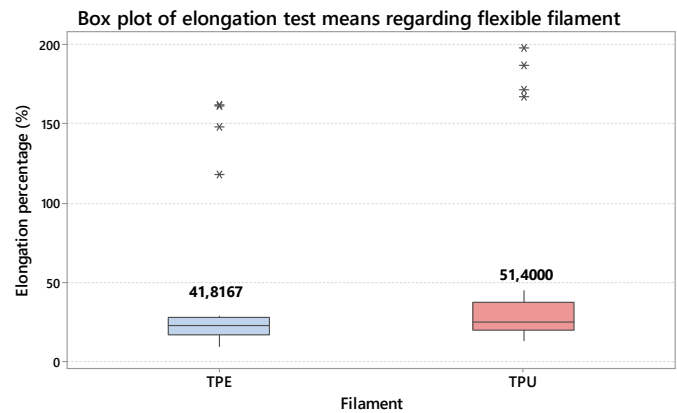


Fig. 33. Box Plot of Elongation Test Means Regarding Flexible Filament.

TABLE XI. DUNCAN'S NEW MULTIPLE RANGE TEST FOR ELONGATION REGARDING PATTERN

Duncan Group	Mean	N	Pattern
A	163.6000	8	Snowflake - Line
B	30.7500	8	Honeycomb - Line
BC	27.5500	8	Honeycomb - Infill
BC	23.3875	8	Flower of life - Line
C	18.3500	8	Snowflake - Infill
C	16.0125	8	Flower of life - Infill

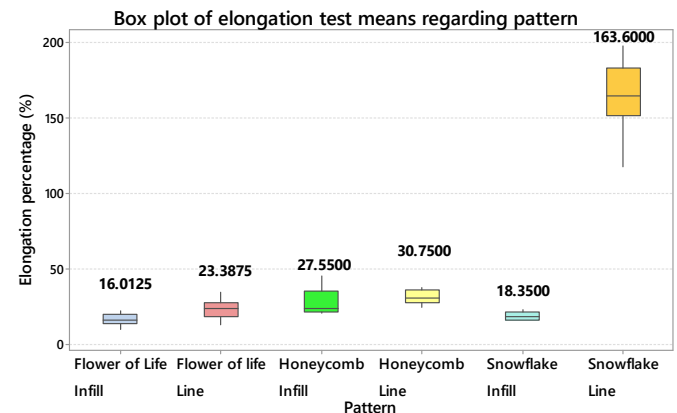


Fig. 34. Box Plot of Elongation Test Means Regarding Pattern.

TABLE XII. RESULTS OF T-TEST FOR THE ELONGATION REGARDING FLEXIBLE FILAMENT

	Levene's Test for Equality of Variances		T-test for Equality of Means					95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	,563	,457	,609	46	,546	9,58333	15,74436	-22,10842	41,27509
Equal variances not assumed			,609	44,238	,546	9,58333	15,74436	-22,14253	41,30920

TABLE XIII. INDEPENDENT SAMPLES FOR ELONGATION TEST REGARDING 3D PRINTER

3D Printer	N	Mean	Std. Deviation	Std. Error Mean
Anet A8	24	46,2417	52,59911	10,73675
M3D Crane Quad	24	46,9750	56,83490	11,60138

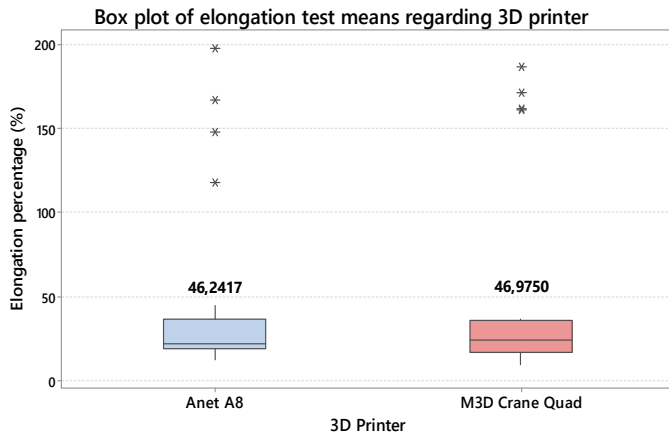


Fig. 35. Box Plot of Elongation Test Means Regarding 3D Printer.

3D Printer. Table XVI show the T-test results ($P \geq 0.710$) for the elongation test among 3D printers. It can be concluded that there is no significant difference between the Anet A8 printer and the M3D Crane Quad printer (see Table XIII and Fig. 35).

B. Tensile

Table XIV shows the data obtained during the tensile test. The test was performed to estimate the amount of force (N) that can be applied to the pattern.

Table XV presents the ANOVA test results, showing that the design is significant and applicable for the two factors: flexible filament and pattern (see fig 36).

Filament factor (A). - The F-value = 15.110 leaves a p-value of 0.000 on the right, less than the 5% significance level. The effectiveness depends on the effect of the filament factor, which is significant; therefore, the design is effective, and the addition is successful. Thus, the tensile test depends on the filament factor when elaborating the garment.

Pattern factor (B). - The F-value = 43.429, leaves a p-value of 0.000, less than the 5% significance level. It means that some pattern significantly influences on the garment design. In other words, there are significant differences between patterns.

A x B. - This row refers to the influence of all interactions between factors. The F-value=0.758 leaves a p-value of 0.586, which is a not significant result. It must be concluded that there is no interaction between filament and pattern.

Error. - This row refers to the variance of the dependent variable not explained by the design.

Total. - This row shows the variance observed in the dependent variable caused by all the factors.

TABLE XIV. DATA OF TENSILE STRENGTH TEST

Filament	Pattern	3D Print			
		A8 Anet		M3D Crane Quad	
		P1	P2	P1	P2
TPU	Snowflake Line	34.70	45.23	29.09	41.20
	Snowflake Infill	50.50	55.36	48.27	54.37
	Flower of Life Line	72.23	79.86	62.62	83.77
	Flower of Life Infill	169.43	140.11	165.67	146.12
	Honeycomb Line	64.40	94.29	28.77	102.13
	Honeycomb Infill	98.28	62.08	61.76	58.54
TPE	Snowflake Line	27.14	31.32	21.34	26.82
	Snowflake Infill	44.94	50.45	38.06	34.44
	Flower of Life Line	42.59	30.18	37.89	30.88
	Flower of Life Infill	141.86	142.03	139.69	139.96
	Honeycomb Line	22.04	76.81	41.35	63.70
	Honeycomb Infill	86.07	18.51	68.50	17.77

TABLE XV. ANOVA RESULTS OF TENSILE STRENGTH TEST

Source	df	Sum of Squares	Mean Square	F	Sig.
Filamento (A)	1	4693.201	4693.201	15.110	0.000
Patron (B)	5	67447.163	13489.433	43.429	0.000
A x B	5	1177.018	235.404	0.758	0.586
Error	36	11181.991	310.611		
Total	47	84499.373			

TABLE XVI. RESULTS OF T-TEST FOR THE ELONGATION TEST REGARDING 3D PRINTER

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	,140	,710	-,046	46	,963	-,73333	15,80727	-32,55171	31,08505
Equal variances not assumed			-,046	45,727	,963	-,73333	15,80727	-32,55684	31,09018

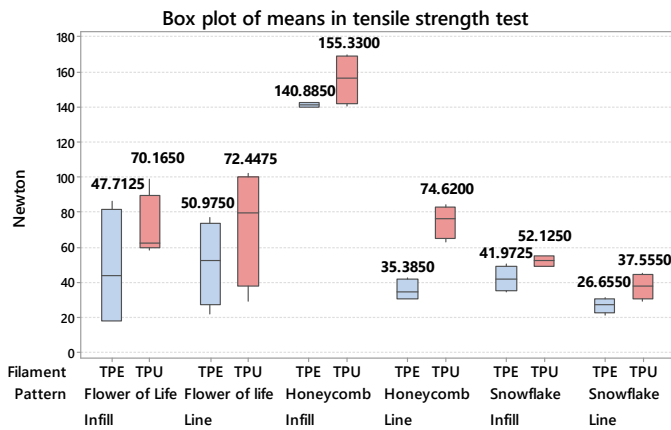


Fig. 36. Box Plot of Tensile Strength Test Means Regarding Pattern and Flexible Filament.

However, the ANOVA does not determine which treatment has the most significant average increase. A T-test was performed between flexible filament (A) and pattern(B) to find the mentioned value.

Flexible Filament (A). Table XIX show the results of the T-test ($P \geq 0.869$). There is no statistical difference between the TPU and TPE filament during the tensile test (see Table XVII and Fig. 37).

TABLE XVII. INDEPENDENT SAMPLES FOR TENSILE TEST REGARDING FLEXIBLE FILAMENT

Filament	N	Mean	Std. Deviation	Std. Error Mean
TPU	24	77,0404	41,17318	8,40444
TPE	24	57,2642	42,12603	8,59894

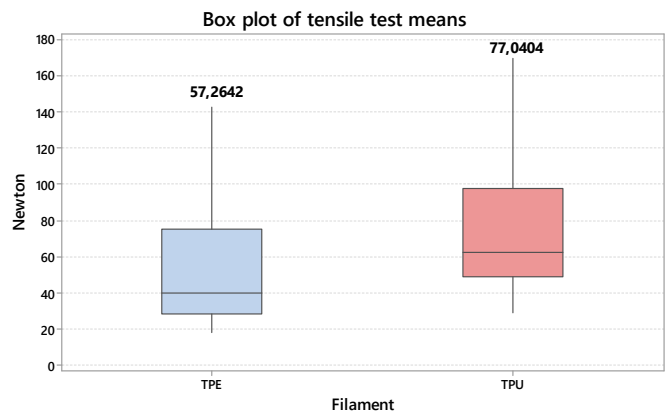


Fig. 37. Box Plot of Tensile Test Means.

Pattern (B). Table XVIII and Fig. 38 show the results of Duncan's new multiple range test ($P \leq 0.05$). It shows exemplars' performance for the Tensile test where the Honeycomb pattern - Infill (148.1075 N.) has a higher average score, indicating that it is statistically different from other designs.

TABLE XVIII. DUNCAN'S NEW MULTIPLE RANGE TEST FOR THE TENSILE TEST REGARDING PATTERN

Duncan Group	Mean	N	Pattern
A	148.1075	8	Honeycomb - Infill
B	61.7113	8	Flower of life - Line
B	58.9387	8	Flower of life - Infill
B	55.0025	8	Honeycomb - Line
BC	47.0487	8	Snowflake - Infill
C	32.1050	8	Snowflake - Line

TABLE XIX. RESULTS OF T-TEST FOR THE TENSILE TEST REGARDING FLEXIBLE FILAMENT

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	,027	,869	1,645	46	,107	19,77625	12,02399	-4,42679	43,97929
Equal variances not assumed			1,645	45,976	,107	19,77625	12,02399	-4,42713	43,97963

TABLE XX. RESULTS OF T-TEST FOR THE TENSILE TEST REGARDING 3D PRINTER

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	,031	,864	-,208	14	,838	-1,34500	6,46151	-15,20356	12,51356
Equal variances not assumed			-,208	13,620	,838	-1,34500	6,46151	-15,23993	12,54993

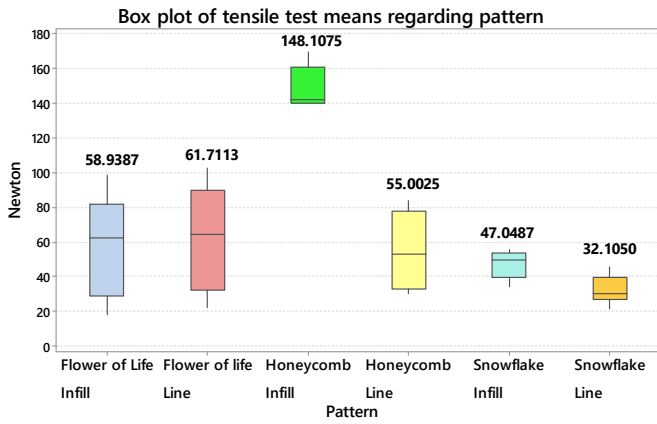


Fig. 38. Box Plot of Tensile Test Means Regarding Pattern.

3D Printer. Table XX show the T-test results ($P \geq 0.864$) for the tensile test among 3D printers. It can be concluded that there is no significant difference between the Anet A8 printer and the M3D Crane Quad printer (see Table XXI and Fig. 39).

TABLE XXI. INDEPENDENT SAMPLES FOR TENSILE TEST REGARDING 3D PRINTER

3D Printer	N	Mean	Std. Deviation	Std. Error Mean
Anet A8	24	66,5575	44.82736	9.15035
M3D Crane Quad	24	67,7471	40.78793	8.32580

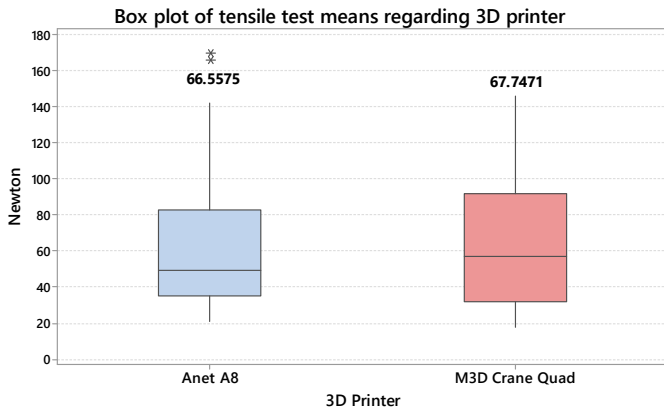


Fig. 39. Box Plot of Tensile Test Means Regarding 3D Printer.

VI. DISCUSSIONS

The elongation and traction test were applied to the natural pattern and not to the garment composed of 29 pieces. It implies that the percentage of extending (elongation) and force applied to the garment itself until reaching the breaking point (tensile test) will have a higher performance.

The elongation test had greater importance when selecting the pattern to elaborate clothing since the elasticity property helps the garment adapt to different sizes (see Fig. 28, 29, 30, and 31).

In the thesis 3D Print Fashion [9], the author was inspired by a basic geometric figure pattern, among which octagons, hexagons, and triangles highlight. Although these patterns

present characteristics similar to those of a mirror, the work does not evidence mechanical tests of tensile or elongation.

Danit [7] was inspired by Eugene Delacroix's "Liberty Leading The People" to select triangles' abstraction as a primary pattern. In the following experiments, Danit used Andreas Bastian's [30] cellular structure pattern; however, there is also no evidence of mechanical testing in his work.

In the work of 3D Print Fashion [9], the creation of a garment by digitalized fabrication of 3D printed surfaces was performed, which demonstrated that it is possible to create textile products with 3D printing technology. Besides, the mechanical design made possible the personalization of products. Such fact introduced the term "Customization" to the textile area.

On the other hand, Danit Peleg experimented with 3D printing, developing 3D manufactured garments. In her proposal for The Shenkar College of Engineering and Design in Israel, she demonstrated that it is possible to modify clothing after its initial design. Afterward, she designed complete collections and, by 2015, she began experimenting with customization [31].

VII. CONCLUSIONS

Concerning the natural patterns proposed for this research and the percentage of elasticity, it is observed that there is significant variation among the results obtained in the elongation test. The Snowflake-line pattern is located in first place with an average value of 163.6% compared to the other designs. In Duncan's group, it was classified in section "A," with a higher value than the patterns organized in other sectors. This test's measurement is important because it states that the garment has the property of adaptability in the different sizes of women's clothing. However, the use of other patterns, especially line patterns, is not discouraged for the textile industry.

The tensile test shows that the honeycomb pattern with infill obtained the best performance (148.1075 N). However, the difference between the design with the best performance and the pattern with the lowest score is minor compared to the elongation test. In addition, data show that the patterns with infill variant achieved a better result since elasticity, which is an essential factor for the elaboration of the garment, has been removed.

Clothing printed with the M3D Crane Quad printer has slightly more elasticity. This statement can be observed in Fig. 29 and Fig. 31 since the garments are larger than those printed with the Anet A8 3D printer.

Lastly, with the flexible filament factor, no significant differences were observed when elaborating the garment's tensile tests. This fact presents that both filaments can be used for the elaboration of clothing.

VIII. FUTURE WORK

It is necessary to research the filament factor in 3D Printing, especially distinguishing the textile industry's flexible and rigid filaments since there is insufficient study. This proposal will help strengthen the line of research in

Technology 4.0, Digital Manufacturing, and 3D Printing as a new textile research trend with the single purpose of developing textile products with filaments.

It is also essential to produce research that increases and improves 3D printing as an innovative knowledge. This approach can be conducted using other factors; filaments of different brands, time, cost, temperature, environment, climate, et al.

Lastly, more research is needed on clothing customization in 3D printing and pattern design, adding adaptability and comfort factors.

REFERENCES

- [1] V. Betancur and F. Zuleta, 'Creación de una prenda mediante la fabricación digitalizada de superficies impresas en 3D', *Iconofacto*, vol. 13, no. 20, pp. 194–206, 2017, doi: 10.18566/iconofacto.v13.n20.a11.
- [2] J. Pérez De Lama Halcón, M. Gutiérrez De Rueda García, JM. Sánchez-Laulhé Sánchez De Cos, and JJ. Olmo Bordallo, 'Fabricación digital, código abierto e innovación distribuida', 2012, [Online]. Available: <http://hdl.handle.net/10251/15018>.
- [3] B. Mandelbrot, *La geometría fractal de la naturaleza*. 2014.
- [4] D. Peleg, '3D Printed Fashion', 2017. <https://danitpeleg.com/> (accessed Aug. 10, 2019).
- [5] A. Jorquera, P. Coronel Romero, and P. I. Warren Alonso, *Fabricación digital introducción al modelado e impresión 3D*. Madrid: Ministerio de Educación, Cultura y Deporte, 2018.
- [6] UNAJ, 'Líneas de Investigación de la Universidad Nacional de Juliaca', 2019. <http://unaj.edu.pe/web2/?p=5904> (accessed Aug. 12, 2019).
- [7] D. Peleg, '3D Printed Jacket', *Danit Peleg*. <https://danitpeleg.com/product/create-your-own-3d-printed-jacket/> (accessed Jan. 30, 2020).
- [8] D. Peleg, *Fashion + 3D Printing*, 2014. <http://danitpeleg.com> (accessed Sep. 06, 2019).
- [9] V. Betancur Fernández, '3D Print Fashion: Creación de una prenda vestimentaria mediante la fabricación digitalizada de superficies impresas en 3D.', Tesis de pregrado, Universidad Pontificia Bolivariana, Medellín, Colombia, 2016.
- [10] W. Wang, G. Zhang, L. Yang, and W. Wang, 'Research on garment pattern design based on fractal graphics', *EURASIP J. Image Video Process.*, vol. 2019, no. 1, p. 29, Dec. 2019, doi: 10.1186/s13640-019-0431-x.
- [11] A. D. K.-T. Lam, 'A study on fractal patterns for the textile design of the fashion design', in *2017 International Conference on Applied System Innovation (ICASI)*, Sapporo, Japan, May 2017, pp. 676–678, doi: 10.1109/ICASI.2017.7988605.
- [12] D. Halápi, S. E. Kovács, Z. Bodnár, Á. B. Palotás, and L. Varga, 'Tensile analysis of 3D printer filaments', presented at the MultiScience - XXXII. microCAD International Multidisciplinary Scientific Conference, 2018, doi: 10.26649/musci.2018.013.
- [13] D. Stechina, S. M. Mendoza, H. D. Martín, C. N. Maggi, and M. T. Piovan, 'Determinación de propiedades elásticas de piezas poliméricas construidas por impresión 3D, sometidas a flexión', *Matér. Rio Jan.*, vol. 25, no. 2, p. e-12617, 2020, doi: 10.1590/s1517-707620200002.1017.
- [14] L. Alvarez C, R. F. Lagos C, and M. Aizpun, 'Influencia del porcentaje de relleno en la resistencia mecánica en impresión 3D, por medio del método de Modelado por Deposición Fundida (FDM)', *Ingeniare Rev. Chil. Ing.*, vol. 24, no. Especial, pp. 17–24, Aug. 2016, doi: 10.4067/S0718-33052016000500003.
- [15] N. G. Tanikella, B. Wittbrodt, and J. M. Pearce, 'Tensile strength of commercial polymer materials for fused filament fabrication 3D printing', *Addit. Manuf.*, vol. 15, pp. 40–47, May 2017, doi: 10.1016/j.addma.2017.03.005.
- [16] D. Chakerian and B. B. Mandelbrot, 'The Fractal Geometry of Nature', *Coll. Math. J.*, vol. 15, no. 2, p. 175, Mar. 1984, doi: 10.2307/2686529.
- [17] P. Valdés, 'Introducción a la geometría fractal', Tesis de pregrado, Universidad del Bío-Bío, Chillán, Chile, 2016.
- [18] A. I. Ramírez-Galarza, *Geometría analítica: una introducción a la geometría*. México: UNAM, Facultad de Ciencias, Coordinación de Servicios Editoriales, 2013.
- [19] J. Garcá, *Fractales*. Salamanca: Facultad de Bellas Artes, 2014.
- [20] B. B. Mandelbrot and J. Llosa, *La Geometría fractal de la naturaleza*. Barcelona: Tusquets, 2003.
- [21] V. Arguedas T., 'La Geometría de la Naturaleza: Benoit Mandelbrot', *Rev. Digit. Matemática Educ. E Internet*, vol. 12, no. 1, Mar. 2014, doi: 10.18845/rdmei.v12i1.1685.
- [22] A. Amadio, 'La Naturaleza Fractal', *Rev. Argent. Psicopedag. ISSN 1514-5603 N° 58 2004*, Jan. 2004.
- [23] W. D. Browning, C. O. Ryan, and J. O. Clancy, *14 Patrones de diseño biofilico*. New York: Terrapin Bright Green, LLC, 2017.
- [24] P. Munro, 'Fractals For Fashion - Textile Weaving Designs', *Fibre2fashion*, 2010. <http://www.fibre2fashion.com/industry-article/4999/fractals-for-fashion-textile-weaving-designs> (accessed Aug. 20, 2020).
- [25] F. Hernandez, P. Palominos, M. Orellana, and M. Ochoa, 'Fractals and their contribution to the design of textile prints', vol. 109, pp. 47–57, Jan. 1996.
- [26] J. F. Francolí and R. B. Díaz, 'Estado actual y perspectivas de la impresión en 3D', *Artíc. Econ. Ind.*, p. 15, 2014.
- [27] F. Bordignon, A. A. Iglesias, and Á. Hahn, *Diseño e impresión de objetos 3D: una guía de apoyo a escuelas*, 1ra. ed. Buenos Aires: Universitaria, 2018.
- [28] F. Leyton, 'Estudio y caracterización de las variables que afectan a la impresión 3D en la generación de objetos manipulables', Investigación, Universidad de la República, Montevideo, Uruguay, 2016.
- [29] O. A. Herrera and M. Figueroa, 'Impresión 3d de proyectos de ingeniería y construcción', Tesis de pregrado, Universidad Andrés Bello, Santiago, Chile, 2017.
- [30] A. Bastian, 'Mesostructured Cellular Materials: Early Prototypes', *Makerbot Thingiverse*, 2014. <https://www.thingiverse.com/thing:289650> (accessed Jan. 30, 2020).
- [31] 'Danit Peleg', *Wikipedia, la enciclopedia libre*. Dec. 20, 2020, Accessed: Feb. 20, 2021. [Online]. Available: https://es.wikipedia.org/w/index.php?title=Danit_Peleg&oldid=131836299.

Smartphone-based Recognition of Human Activities using Shallow Machine Learning

Maha Mohammed Alhumayyani¹
Information Systems Department
Faculty of Computer and
Information Sciences
Ain Shams University, Egypt

Dr. Mahmoud Mounir²
Faculty of Computer and
Information Sciences
Ain Shams University, Egypt

Prof. Dr. Rasha Ismael³
Vice-dean fea graduate studies and
research, Faculty of Computer and
Information Sciences
Ain Shams University, Egypt

Abstract—The human action recognition (HAR) attempts to classify the activities of individuals and the environment through a collection of observations. HAR research is focused on many applications, such as video surveillance, healthcare and human computer interactions. Many problems can deteriorate the performance of human recognition systems. Firstly, the development of a light-weight and reliable smartphone system to classify human activities and reduce labelling and labelling time; secondly, the features derived must generalise multiple variations to address the challenges of action detection, including individual appearances, viewpoints and histories. In addition, the relevant classification should be guaranteed by those features. In this paper, a model was proposed to reliably detect the type of physical activity conducted by the user using the phone's sensors. This includes review of the existing research solutions, how they can be strengthened, and a new approach to solve the problem. The Stochastic Gradient Descent (SGD) decreases the computational strain to accelerate trade iterations at a lower rate. SGD leads to J48 performance enhancement. Furthermore, a human activity recognition dataset based on smartphone sensors are used to validate the proposed solution. The findings showed that the proposed model was superior.

Keywords—Data preprocessing; data mining; classification; genetic programming; Naïve Bayes; decision tree

I. INTRODUCTION

The aim of human action recognition (HAR) is to recognize activities extracted from a number of observations concerning the behavior and environmental conditions of subjects. A lot of applications for HAR research include video monitoring, healthcare and contact with human-computer. HAR uses sensors influenced by human movement for the classification of an operation of the individual. Both users and sensors of smartphones expand as users also bring their smartphones. HAR seeks to identify activities arising from a variety of observations concerning the behavior and environmental conditions of subjects.

Sensors can help patients always record and track and automatically report if abnormal behavior has been detected by a huge quantity of resources. The research benefits from other applications, including the human survey method and position predictor. Many experiments have successfully established wearable sensors with a low error rate, but most work is conducted in labs with very limited settings. Readings from

many body sensors achieve a low error rate, but in reality the complex environment cannot be achieved [1].

The efficiency of the human action mechanism can be deteriorated by several challenges. One is that the extracted features need to generalize many variations in order to address the challenges of action recognition, including individual appearances, viewpoints and histories. In addition, the relevant classification should be guaranteed by those features. The creation of a lightweight, precise device on Smartphones that can detect human activities and reduce labelling time and burden is another challenge.

The main purpose of this paper is to reliably detect the type of physical activity that the user conducts using the phone sensors. This involves an analysis of existing solutions, finding ways to strengthen them and finding a new approach to the issue. Furthermore, a human activity recognition dataset based on smartphone sensors are used to validate the proposed solution. Section 2 is associated work on recent study events in the field of methods and applications for human action detection. Section 3 describes the basic methodologies and principles. Section 4 addresses with shallow learning the proposed method of human behavior recognition. Section 5 symbolizes the findings of the experiment. The conclusion of Section 6 is the representation of the result of the proposed scheme.

II. RELATED WORK

Anguita et al. [2] introduced a system that uses inertial smartphone sensors to recognise human physical activity (AR). Since the energy and computer power of these mobile phones is small, they suggest a new hardware-friendly method for classification of multi-class problem. This approach adapts the regular Support Vector Machine (SVM) and uses fixed-point arithmetic for the reduction of computational costs.

Tran and Phan [3] have created and built a smartphone framework for human activities through the use of integrated sensors. For acknowledgement, six acts are selected: standing, upstairs, walking, sitting, downstairs, lying down. The Support Vector Machine (SVM) for classification and identification of the operation is used in this method. For the model classification model - the model file, data obtained from sensors is analyzed. The classification models are optimized to generate the best results for the human activity described.

In the sense of human recognition of human behavior, Gusain et al. [1] evaluated gradient boosted machines (GBM). The proposal solution uses an ensemble of SVM to incorporate incremental learning. After the first batch of data has been trained, the computer is stored in many machines. This machine is trained on the new batch for the second time, and correctly classified information is removed, but the misclassified machine is trained.

The generic feature engineering approach Zdravevski et al. [4] have proposed to pick robust characteristics from a variety of sensors that can be used to generate accurate classification models. A number of time and frequency domain features have been extracted in the initially registered time series and some newly created time series [i.e. fast Fourier transformation series, first derivatives, magnitudes and Delta series]. Also, the number of functions generated is substantially reduced with a two-phase function selection. Finally, various classification models are trained and tested in a separate test collection.

Hassan et al. [5] proposed an inertial smartphone sensor method for detection of human behavior. Second, raw data extract productive functionality. The characteristics include autoregressive coefficients, meaning, median, etc. A Linear Discriminant Analysis (LDA) and kernel principal component analysis (KPCA) further process the features to make them robust. The features are eventually educated in effective

identification of behavior with the Deep Belief Network (DBN).

Xu et al. [6] proposed an InnoHAR deep learning model based on a neural network and a recurring neural network. The model enters end-to-end multi-channel sensor waveform data. Multi-dimensional functions with different kernel-based convolution layers are extracted in initial modules. In conjunction with GRU, time series characteristics are modelled and data characteristics are used entirely to complete classification tasks.

Inertial smartphone accelerometer architecture design for HAR has been developed by Wan et al. [7]. In traditional everyday activities, the smartphone gathers the sensory data sequence and extracts from the original data the high efficiencies, and then uses several three-axis accelerometers to acquire the physical behavioral data of the consumer. The data are preprocessed to extract useful feature vectors by denoising, normalizing and segmenting. A real-time method of classification of human behavior based on the neural convolution network (CNN) using a CNN for the extraction of local functions is also suggested.

Next, Table I compares between recent literatures of different researchers works on human activity recognition. The table spots the major drawbacks and dataset(s) already used for benchmarking.

TABLE I. SUMMARY OF RECENT LITERATURE REVIEWS

Author	Year	Dataset	Technique	Results	Advantages	Drawbacks
Anguita et al. [2]	2012	Human Activity Recognition	Adapted support vector machine (MC-HF-SVM)	Accuracy (89%)	Improvement in computational cost	The same accuracy as the standard SVM
Tran and Phan [3]	2016	Human Activity Recognition	Support vector machine	Accuracy (89.59%)	A full human activity recognition system has been created. The system completed included systems for data collection, retrieval, analysis, training and identification of human activities. This system can be used in many fields of practice, in particular in the field of health care.	There are also a few restrictions in the system. Some indicators have a low percentage of acceptance.
Gusain et al. [1]	2017	Human Activity Recognition	extreme Gradient Boosted Decision Trees	Accuracy (90%)	The proposed solution is more accurate than the conventional approach and requires far less time. The findings are positive and show the truthfulness of the model.	More complexity in the system and need real-time processing
Zdravevski et al. [4]	2017	Human Activity Recognition and other four datasets	6 classifiers	accuracy (95.8%)	Identified that a high degree of activity recognition can be achieved with just a smartphone and smart watch. This result is	Data imbalance issue in applications for identification of human behaviors in real life.

					important and inspiring since the two devices are now widespread and not too invasive.	
Hassan et al. [5]	2018	Human Activity Recognition	(KPCA) and (LDA) +a Deep Belief Network (DBN)	Accuracy (95.85%)	The solution proposed was compared to the conventional SVM multi-class approach where its superiority was seen. It has also demonstrated its ability to differentiate between fundamental transitional and non-transitional practices.	More complexity in the system and need real-time processing
Xu et al. [6]	2019	Opportunity activity recognition	InnoHAR	Accuracy (94.6%)	The proposed approach has consistently superior performance in three most commonly used public datasets and strong generalization performance. During the experiment, they also showed that in real time, practices on the Minnow Board Turbot Dual Core Board have more potential with its creative structure.	Data imbalance issue in applications for identification of human activities in real life.
		PAMAP2		Accuracy (93.5%)		
		Human Activity Recognition		Accuracy (94.5%)		
Wan et al. [7]	2020	Human Activity Recognition	CNN	Accuracy (92.71%)	The advantages and drawbacks of human enforcement are contrasted with five algorithms, CNN, MLP, BLSTM, SVM and LSTM. Experiments show that CNN still has considerable value in recognition of human behavior and is an outstanding classification and recognition algorithm.	Further optimization is possible in the structure of the four neural network models used in the experiment and further comparative studies can be carried out.
		Pamap2		Accuracy (91%)		

III. PRELIMINARIES METHODS

A. Shallow Learning

Machine learning is seen as a form of artificial intelligence (AI) which deliver learning-free machines with no more processes and Shallow learning [8] is regarded as machine learning. They have evolved from theory of machine learning and pattern recognition. Two key categories of learning are typically un-supervised and supervised. The training set comprises samples of input vectors and matched objective vectors for supervised learning. No labels are required for the training set in unsupervised training. The supervised target of learning is to predict an adequate output vector for each vector. Classification tasks are functions where the objective label is a discrete finite number of the group. It is difficult to describe the unsupervised learning target. The related samples of sensitive clusters within input data, known as clustering, are a primary objective.

B. Genetic Programming (GP)

Genetic programming (GP) is a technique of evolutionary computing (EC) that solves problems automatically without asking the machine how to do this [9] directly. From the most abstract level, GP is a domain-independent, systematic way to get computers to solve problems automatically, from a high-level argument. GP is a special evolutionary algorithm (EA) where computer programs are present in the population. GP thus converts populations of programs, as shown in Fig. 1, from generation to generation [10].

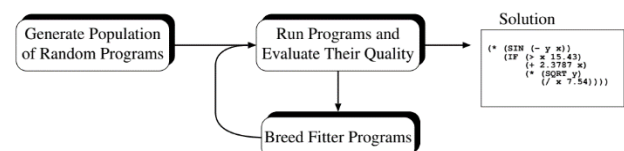


Fig. 1. Block Diagram of Genetic Programming.

Any computer application, with ordered branches, can be graphically displayed as a rooted label tree. Genetic programming is an enhancement to the conventional genetic algorithm, which is a computer programme for every person in the population. The genetic programming search area is the space for all possible computer applications, which consist of functions and terminals that are suitable for the problem area. The features may include standard arithmetic operations, logical functions, standard programming, standard math, or domain-specific functions.

Five preparatory steps have been taken [11]. The following five steps are:

- 1) The terminal package,
- 2) The elementary functions set,
- 3) The measure of fitness,
- 4) The run control parameters, and
- 5) The formula for the outcome and the end of the run criterion.

In preparation for genetic programming the first step is to classify the terminal sequence. The terminals can be seen as the entries into the computer program that has been uncovered. Terminals are the ingredients from which genetic programming tries to construct or approximately solve a computer program to solve the problem.

The second step in planning for the use of genetic programming is to recognize the set of functions to produce the mathematical expression to match the unique finite data sample. The functions of the F function set and the terminals of the T terminal set are used in any computer program. In each function set, any value and data type that may be returned to a function set and to any value and data type that may be assumed by the terminal in the set should be recognized as its arguments. That is, the selected function set and terminal set should be closed. These first two steps correspond to the step of specifying the representation scheme for the conventional genetic algorithm. These two first steps correspond to the step of defining the traditional genetic algorithm representation scheme.

The remaining genetic programming steps are the three last preparatory steps for typical genetic algorithms. Populations of hundreds, thousands and millions of computer systems are genetically derived from genetic programming. This breeding takes place using the Darwinian survival and reproduction concept of the most suitable and genetic crossover operation for computer-based programming. This combination of Darwinian natural selection and genetic operations frequently results in a computer program that solves a given problem. Genetic programming begins with an initial population of computer programs randomly generated (generation 0) consisting of features and terminals for problem domain applications.

The establishment of this initial random population is essentially a blind random search of the problem's search space as a computer program. Each computer program within the population is calculated by the fitness of the problem area. The fitness calculation is different from the problem [12]. A combination of the number of properly treated instances (i.e.,

true negatives and true positives) and the number of correct instances will calculate the fitness of a program (i.e., false positives and false negatives). Correlation is also used as a test of fitness. From the other hand, the fitness of a particular computer program may be calculated using entropy, the fulfilment of the gap test, the success test satisfaction or a combination of these. For several problems a combination of factors like correctness, parsimony (smallness in the program), or effectiveness (of execution) may be needed to use a multifunctional fitness measure [12].

In general, each computer program in the population has numerous different fitness instances, with the consequence that it is evaluated in a number of representative situations, either in total or in average. These fitness instances may be a sampling of different independent variable values or a sampling of different initial system conditions. The fitness cases can be picked alone or arranged (e.g., over a regular grid or at regular intervals). The initial conditions are also usual in cases of fitness (as in a control problem). In genetic programming generation 0, computer programs are almost always extremely suited.

Some people would however prove more suited than others in the population. These output disparities are then taken advantage of. A new descendant population of individual computer programs is being generated through the Darwinian theory of reproduction and survival of the fittest and genetic fusion process. The reproductive method involves choosing a computer program, which can be used by copying into new population [13], from the existing population of fitness-based programs. The crossover operation is used to construct new descending computer programs from two fitness-based parental programs.

The genetic programming parental systems are of various sizes and types. The offspring programs consist of their parents' sub-expressions (building block, sub-programs, subtrees, subroutines). These descent programs, as opposed to their parents, have different sizes and styles. In genetic programming the mutation operation can also be used. The population of offspring (i.e., the new generation) replaces the old population after the genetic operations on the present population (i.e., the old generation). Each participant in the new program population will then be assessed for fitness and over several generations the process is replicated. At every point, the state of the process will consist only of the present population of people in this highly parallel, locally regulated, decentralized process. The driving force behind this mechanism is just the human health in the existing population that has been observed. As can be seen from this algorithm, populations of programs are generated that appear to display an increasing average fitness in their environment over several generations. Furthermore, these machine populations are able to adapt quickly and efficiently to environmental changes. The best person in any run is usually referred to as the outcome of the course of genetic programming [13]. Inherently hierarchical are the products of genetic programming.

Sometimes, genetic programming effects are default hierarchies, priority hierarchies of tasks, or hierarchies where one action subordinates or suppresses another. Another

fundamental characteristic of genetic programming is the dynamic variability of the computer programs that are built along the way to a solution [10]. The effort to describe or minimize in advance the dimensions and form of the potential solution is always hard and unnatural. Furthermore, advancing or restricting the solution's dimensions and type narrows the window through which the machine sees the world and may prevent the solution of the problem from ever being found. The absence or a relatively minor function of inputs and post-processing inputs is another important aspect of genetic programming. Usually, the inputs, intermediate effects and outputs are directed in the natural terminology of the problem region. Genetic programming systems consist of functions that are natural to the problem area. If appropriate, a wrapper will conduct the post processing of the output of a program (output interface). Eventually, another key element of genetic programming is the active genetic programming structures [11]. They really aren't inactive encodings of the problem (i.e. chromosomes). Instead of running a machine, the genetic program structures are active structures that can be run as they are.

C. Decision Tree

The classification in decision trees [14] is based on a sequence of decision classification of the sample. The present decision helps to make a subsequent decision in a decision tree to create a sequence that is indicative of the structure of the tree. The structure includes two key types of attributes and allows to use attributes during the prediction process. The predicted attribute is described as a dependent variable because the value of the other attributes depends on or depends upon the values. The other attributes that help to forecast the dependent variable value are known as the independent dataset variables. In the case of classification, each end leaf node represents one decision or category, the root node becomes an eligible end leaf node, for instance. Each node has the attributes of the instances and the value of each division is the same as the attributes of each division. The decision tree is a model that determines the value of the dependent variable(s) based on the values of various attributes of the data available in a new case. In decision tree, The inner nodes indicate the various attributes The divisions between the nodes reflect potential values in the observed samples for these attributes, whereas the classification of the dependent variable or the final value are represented by the terminal nodes.

After the related basic calculation, the J48 [15] decision tree classifier shall be used. In order to order anything else, the ultimate objective is to create a selection tree first because of the quality estimates of the accessible preparation information. Therefore, whatever the stage in which things are organized (training), it identifies the characteristic that usually clearly distinguishes the various occurrences. This distinctness, which has the capacity to get the best out of the instances, may be structured to collect the necessary data. At present, if the standard for which the information events falling into its class have the same meaning for the target variable is of some fair value that there is no vagueness, this expansion would be terminated and designated as the objective value that could be achieved. For alternative situations, the search for alternate quality begins which results in the most astonishing data

collected—and goes on until either a fair choice has been made about which combination of the unique characteristics of a particular target quality or the use of properties. If qualities are used or if an exact result from the accessible information cannot be obtained, the degradation of this extension goal was priced for most of the items in this branch. [5]. See Table II which compares between different algorithms can be used for building decision tree.

D. Naïve Bayes Classifier

One of the most popular simple machine learning classifiers is a probabilistic classifier. Due to the use of probability distribution over a set of classes, the classifier can predict a sample instance instead of predicting only one class for the sample. Probabilistic classifiers have a certain description that can be useful when classifiers are combined into ensembles. Naïve Bayes is known by its probabilistic designation as the straightforward street algorithm. Naïve Bayes is a statistical classification that calculates the likelihood of a certain class of tuple based on the Bayes theorem [16]. The class-conditional Independence characterizes Naïve Bays, implying that the influence of an attribute-value on a certain class is irrespective of the other attributes. High accuracy, pace and many advantages are of Naïve Bayes. In principle, in contrast with all other classifiers, minimum error rate is the main characteristic of Bayesian classifiers [17].

Naïve Bayes is working on a basic definition, but a very intuitive one. In certain instances, Naïve Bayes beats several comparatively complex algorithms by using the variables in the data sample and by observing each other separately and independently. The classification of Naïve Bayes is based on the conditional probability rule of Bayes. It begins with all of the attributes in the data that are equally relevant, independent from one another and is evaluated individually. It works with the hypothesis that one feature works without the others in the study. The model offers a response to questions such as "What is the likelihood of a certain type of attack, provided certain device events, when it comes to using Naïve Bayes, NB in model intrusion attacks? The query in turn is reworded in the context of conditional probability. A directed acyclic (DAG) charts the structure of an NB. Each node represents one of the system variables and each relation codes the effect of one node on the next. So A directly influences B when it has a relation from node A to node B [6].

TABLE II. COMPARISON BETWEEN DIFFERENT ALGORITHMS FOR BUILDING DECISION TREES

Characteristics	CART	C4.5	ID3
Pruning	Post pruning	Pre-pruning	Not applicable
Improvement	Accepted	Not accepted	Not accepted
The lack of values	Can handle the issue	Cannot handle the issue	Cannot handle the issue
Data type	Nominal & Continuous	Categorical & Continuous	Categorical
Formulation	Gini index is used	Gain ratio is used	Information gain & entropy are used
Rapidity	On-average	More quickly than ID3	Low speed

E. Proposed Human Action Recognition System (HARS)

1) *Problem formulation:* Recognition of human activity (HAR) is a method of pattern recognition. Since preprocessing, feature engineering and assignment to labels are the key recognition processes, HAR has applied the method to all sub processes mentioned. In order to identify an action as a label, it is important to preprocess the input data and evaluate it in order to detect whether there is an abnormal value...etc. Due to the acquiesced data from sensors of the smartphone, which represents the characteristics and features of human activity. The final step before classification, however, includes the study and engineering of features. We addressed each method for more details in depth in the following sections.

2) *Shallow human action recognition system:* The framework proposed is planned and built in two phases: the first phase consists of the pre-processing of acquired data values, feature engineering and analysis. In contrast, the second stage is the process of using shallow learning algorithms and applies classification based on a Decision tree, Naïve Bayes, and Genetic Programming. This paper proposes

a comparatively shallow learning algorithm for human action recognition based on smartphones. We have taken full advantage of their strengths. Fig. 2 demonstrates our proposed recognition model's abstract architecture. The stream data is pre-processed in instances and is inserted into the classifier in each instance. The shallow learner infers the corresponding action of the input tuple. We have adopted a decision tree architecture to automatically learn efficient and robust action features from the training instances.

Stage (1): Pre-processing

Data is first prepared in the attributes at the last index of the data set class name. In order to minimize the difficulty of the performance assessment of the proposed model, the module data of the class is divided into two key categories. The dataset had already been prepared and made available online. For easy study, activity labels are then changed by having the class label 1, 2, 3, 4, 5 or 6 for WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING or LAYING, respectively. The datasets were transferred with the best search evaluator to the feature analysis with the selection of correlation features to analyze the correlated and uncorrelated features during the processing.

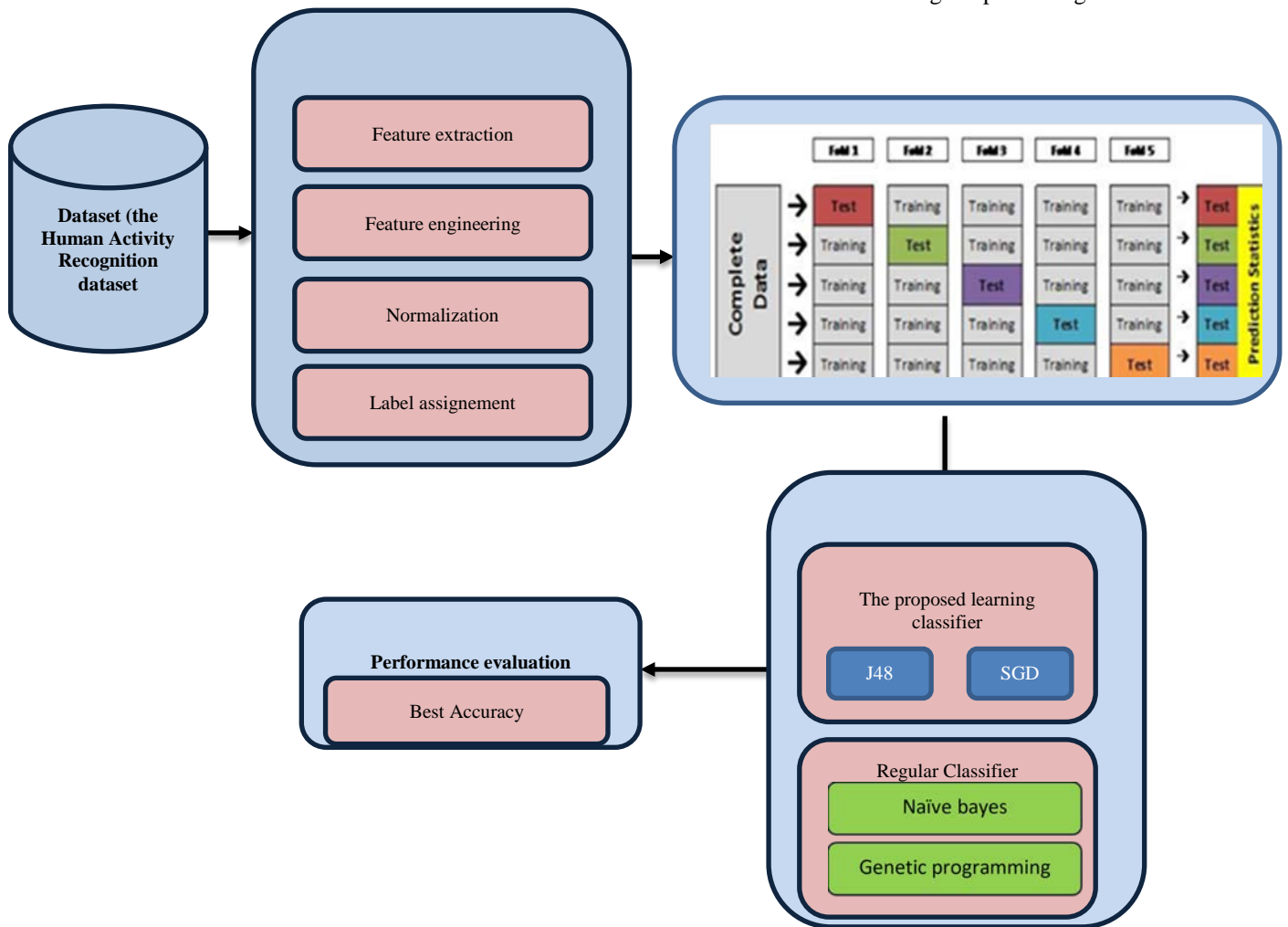


Fig. 2. The Block Diagram for the Proposed Model.

Stage (2): Modeling using Shallow Learning Algorithm

1) *Dataset splitter*: The process of loading data set into two components obtained from the preprocessing module is the process which divides the dataset. The technique of cross-fold validation [18] divides the date set into 2 parts when data is partitioned randomly into k-fold independent set, group is a training set and fold is a test set. The training set is used to train the proposed system while checking and validating the accuracy of the trained model is carried out.

2) *Learning phase*: In the proposed system, the learning phase starts the learning process of the classifier using the current instance from the training dataset that was realized. The results of this base classifier are considered to be the input data for the second stage. The update to the classifier has its own rule set for every detail to be independently conscious of the behavior of human activity.

Stochastic descent [19] (often SGD) is an iterative way to refine an objective function with sufficient smoothness properties (e.g. sub differentiable or differentiable). The gradient optimization can be seen as a stochastic approx. because the real gradient (calculated from the whole dataset) is replaced by an approximation (calculated from a randomly selected subset of the data). This reduces the machine burden, achieving faster iterations in trade at a lower convergence rate, particularly in high-dimensional optimization problems. This is the stochastic gradient descent used to optimize the parameters of j48.

Stage (3): Action recognition stage

The proposed framework uses the test dataset developed by the Data Splitting module during this stage. The test data set is used to assess the model's output. The results of this module are redirected to create the complete classifier performance assessment process that is discussed in detail in the experimental section using the classification rules created during the learning phase for the assessment of the proposed model.

IV. EXPERIMENTAL RESULTS

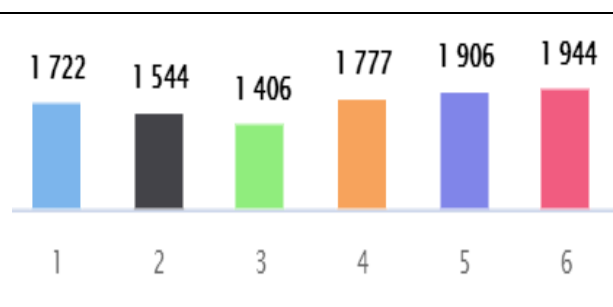
A. Runtime Environment

On a computer system with features the proposed recognition system has been implemented and designed using JAVA 8. Such features have been listed as hardware, comprising 64-bit machine and Intel of Core i7 and 2.2 GHz rather than as a software framework for Windows 10 Professional.

B. Data Set(s)

To evaluate the proposed system, we have used a dataset of Human Activity Recognition provided by Jorge L et al.[2]. It is available online for public use on the UCI repository since 2013. It contains six types of actions, including, LAYING, SITTING, WALKING, STANDING, WALKING_DOWNSTAIRS and WALKING_UPSTAIRS. All has a smartphone on the waist (Samsung Galaxy S II). This dataset version contains all the training and testing examples provided in the original data repository. The data collection was randomly divided into two groups, with 70% of the volunteers chosen to produce the training data and 30% of the test data. The database is composed of the logs of 30 people who conduct everyday life activities (ADL) with an embedded inertial sensor waist-mounted smartphone. In an age group of 19-48 years, the tests were performed with 30 volunteers. We captured three axial linear acceleration and three axial angular speeds at constant speeds of 50Hz using its embedded accelerometer and gyroscope. The tests were captured by video to manually mark the data. The sensor signals (gyroscopes and accelerometers) were pre-processed by using noise filters and sampled in 50 percent overlap and 2.56 sec and (128 readings/windows) in fixed-wide sliding windows. A Butterworth low-pass filter into body acceleration and gravity separated the sensor acceleration signal with gravitational and body movement elements. It is believed that the gravitational table force is only low frequency; a filter with a cutoff frequency of 0,3 Hz has thus been utilized. The measurement of time and frequency domain variables obtained a vector of features from each window. Table III represents a summary of these dataset.

TABLE III. SUMMARY OF THE DATASET

Item	Description
Number of Instances	10299 
Number of features	562
Number of Labels	6
Has missing values	No

V. PERFORMANCE MEASURE

In order to test the accuracy of the classification, the Classifier Performance Evaluator [20] applies various classification performance tests. These measurements are based on certain concepts such as TN (True Negative), FN (False Negative), TP (True Positive) and FP (False Positive).

In Table IV, a table visualization of algorithmic output is represented by the confusion matrix.

TABLE IV. PERFORMANCE MEASURE

Metric	Formula	Metric	Formula
accuracy	$\frac{TP + TN}{P + N}$	Recall, sensitivity	$\frac{TP}{P}$
Precision	$\frac{TP}{TP + FP}$	error rate	$\frac{FP + FN}{P + N}$
Specificity	$\frac{TN}{N}$		

Precision is the most sensitive and interesting measure to be used to compare the fundamental classifiers and the proposed system in a detailed range.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The results of the proposed model are provided in this section. The results provided a comparison between the proposed as an application of three main classifiers. According to Table V, the J48 achieved better accuracy than the naïve Bayes and the genetic programming. These results displayed graphically in Fig. 3. The J48 take the advantages of no domain knowledge required, no parameter setting, can handle multidimensional data, simple and fast.

TABLE V. ACCURACY EVALUATION OF THE PROPOSED CLASSIFIERS

Classifier Algorithm	Accuracy
Decision Tree (J48)	96.6387
Naïve Bayes	94.958
Genetic Programming	94.958

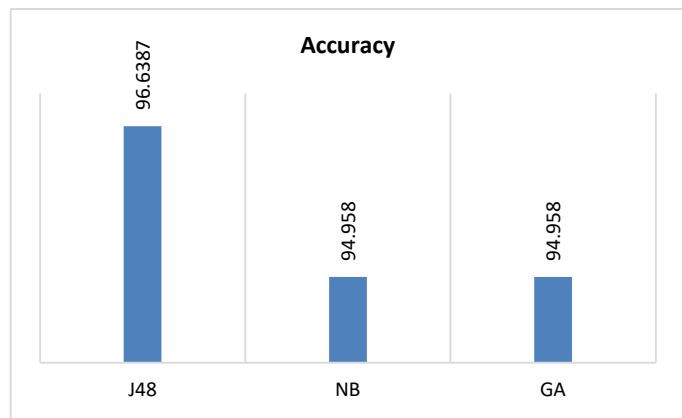


Fig. 3. The Proposed Model Results.

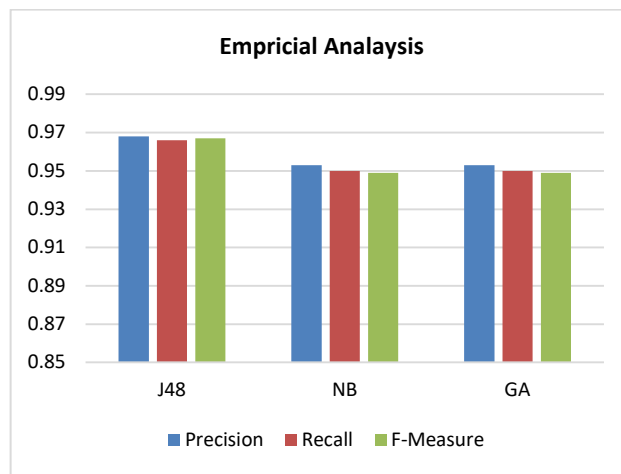


Fig. 4. Empirical Analysis of the Proposed Optimized Algorithm J48 with SGD versus Naïve Bayes and Genetic Classifiers.

Fig. 4 compares the major enhancements of the proposed optimized classifier J48 with SGD versus other regular classifiers including Naïve Bayes and genetic classifiers. The proposed classifier achieves the highest values in precision, recall and F-score measure. While other classifiers almost have the same level of achievements regards the different metrics.

TABLE VI. A COMPARISON TO THE LITERATURE

Authors	Algorithm	Accuracy
Anguita et al. [2]	Adaptive SVM	89
Tran and Phan [3]	SVM	89.5
Gusain et al. [1]	Extreme Gradient DT	90
Hassan et al. [5]	DBN	95.85
Xu et al. [6]	InnoHAR	94.5
Wan et al. [7]	CNN	92.7
The Optimized Proposed Classifier	Hybrid J48-SGD	96.6387

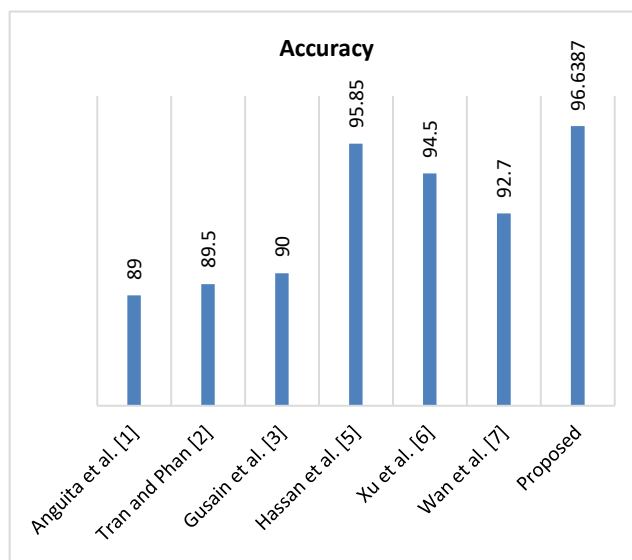


Fig. 5. A Comparison to the Literature Regards to Accuracy Metric.

Moreover, the proposed model compared to the literature according to the accuracy parameter. The results accepted the supremacy of the proposed model, as seen in Table VI and Fig. 5, in achieving the best recognition rate rather than the sixth compared models.

VII. CONCLUSION

If sensors help patients register and monitor them all time and automatically report because of any abnormal activity is found, a massive number of resources may be saved. This paper presented a model to detect how effectively the user conducts physical activity using the sensors of the telephone. The model utilized three different classifiers to test the percentage of the recognition. The j48 with stochastic gradient descent approved its superiority rather than naïve Bayes and genetic programming. Moreover, the model compared to the literature and success to achieve the highest accuracy reached to 96.6%. In future work, the proposed model will be integrated to different optimization algorithm like the slap swarm optimization algorithm, the crow search algorithm or the grey wolf optimization algorithm to improve the recognition rate.

REFERENCES

- [1] K. Gusain, A. Gupta, and B. Popli, "Transition-aware human activity recognition using extreme gradient boosted decision trees," in *Advanced Computing and Communication Technologies*, Springer, 2018, pp. 41–49.
- [2] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*, 2012, pp. 216–223.
- [3] D. N. Tran and D. D. Phan, "Human activities recognition in android smartphone using support vector machine," in *2016 7th international conference on intelligent systems, modelling and simulation (isms)*, 2016, pp. 64–68.
- [4] E. Zdravevski et al., "Improving activity recognition accuracy in ambient-assisted living systems by automated feature engineering," *Ieee Access*, vol. 5, pp. 5262–5280, 2017.
- [5] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Futur. Gener. Comput. Syst.*, vol. 81, pp. 307–313, 2018.
- [6] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "InnoHAR: a deep neural network for complex human activity recognition," *Ieee Access*, vol. 7, pp. 9893–9902, 2019.
- [7] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mob. Networks Appl.*, vol. 25, no. 2, pp. 743–755, 2020.
- [8] Q. Tian, J. Li, and H. Liu, "A method for guaranteeing wireless communication based on a combination of deep and shallow learning," *IEEE Access*, vol. 7, pp. 38688–38695, 2019.
- [9] S. Nguyen, Y. Mei, and M. Zhang, "Genetic programming for production scheduling: a survey with a unified framework," *Complex Intell. Syst.*, vol. 3, no. 1, pp. 41–66, Mar. 2017.
- [10] M. Sukanuma, S. Shirakawa, and T. Nagao, "A genetic programming approach to designing convolutional neural network architectures," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2017, pp. 497–504.
- [11] A. H. Gandomi, A. H. Alavi, and C. Ryan, Eds., *Handbook of Genetic Programming Applications*. Cham: Springer International Publishing, 2015.
- [12] K. Nag, T. Pal, and N. R. Pal, "ASMIGA: An Archive-Based Steady-State Micro Genetic Algorithm," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 40–52, Jan. 2015.
- [13] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Comput.*, vol. 8, no. 1, pp. 3–15, Mar. 2016.
- [14] M. Li, H. Xu, and Y. Deng, "Evidential decision tree based on belief entropy," *Entropy*, vol. 21, no. 9, p. 897, 2019.
- [15] D. Bienvenido-Huertas, J. E. Nieto-Julián, J. J. Moyano, J. M. Macías-Bernal, and J. Castro, "Implementing artificial intelligence in H-BIM using the J48 algorithm to manage historic buildings," *Int. J. Archit. Herit.*, pp. 1–13, 2019.
- [16] I. M. El-hasnony and O. H. Al-tarawneh, "A Proposed Hybrid Effective Technique for Enhancing Classification Accuracy," no. December, 2017.
- [17] Y.-C. Zhang and L. Sakhanenko, "The naive Bayes classifier for functional data," *Stat. Probab. Lett.*, vol. 152, pp. 137–146, 2019.
- [18] I. M. El-Hasnony, S. Baracat, M. Elhoseny, and R. R. Mostafa, "Improved Feature Selection Model for Big Data Analytics," *IEEE Access*, 2020.
- [19] R. Sweke et al., "Stochastic gradient descent for hybrid quantum-classical optimization," *Quantum*, vol. 4, p. 314, 2020.
- [20] S. Alemam, S. Abuelsadat, S. Saber, and T. Elsewify, "Accuracy, sensitivity and specificity of three imaging modalities in detection of separated intracanal instruments," *G. Ital. Endod.*, vol. 34, no. 1, 2020.

Deep Learning Approaches for Intrusion Detection in IIoT Networks – Opportunities and Future Directions

Thavavel Vaiyapuri¹, Zohra Sbai², Haya Alaskar³, Nourah Ali Alaseem⁴

College of Computer Engineering and Sciences,
Prince Sattam bin Abdulaziz University, Saudi Arabia^{1,2,3,4}
National Engineering School of Tunis, Tunis El Manar University, Tunisia²

Abstract—In recent years, the Industrial Internet of things (IIoT) is a fastest advancing innovative technology with a potential to digitize and interconnect many industries for huge business opportunities and development of global GDP. IIoT is used in diverse range of industries such as manufacturing, logistics, transportation, oil and gas, mining and metals, energy utilities and aviation. Although IIoT provides promising opportunities for the development of different industrial applications, they are prone to cyberattacks and demands for higher security requirements. The enormous number of sensors present in the IIoT network generates a large amount of data and has attracted the attention of cybercriminals across globe. The intrusion detection system (IDS) that monitors the network traffic and detects the behaviour of the network is considered as one of the key security solution for securing IIoT application from attacks. Recently, the application of machine and deep learning techniques have proved to mitigate multiple security threats and enhance the performance of intrusion detection. In this paper, we present a survey of deep learning-based IDS technique for IIoT. The main objective of this research is to provide the various deep learning-based IDS detection methods, datasets and comparative analysis. Finally, this research aims to identify the limitations and challenges of existing studies, solutions and future directions.

Keywords—Industrial Control System; Industrial Internet of Things (IIoT); cybersecurity; intrusion detection system and deep learning

I. INTRODUCTION

Industrial Internet of Things (IIoT) provides all industries with more excellent connectivity that, in turn, creates valuable information and intelligence regarding operations. IIoT is utilized in diverse range of industries to interface information, service, and people for intelligent operations in multiple management domains such as smart power, smart city, healthcare, automation industry, agriculture, logistics and transportation [1]. It connects with various sensors and maintains the critical infrastructure, and it requires a larger network scale. IIoT aims to provide intelligent manufacturing goods that establish smart factories with efficient communication between business partners and customers [2]. Industry 4.0 focused on the optimization problems in the industry to utilize data-driven services by using smart devices. This intelligence can achieve more efficiencies and manufacturing enhancements. However, this expanded network opens up these connected devices to specific threats of cyber-attacks. The industrial facilities become more connected, and hackers are more sophisticated. Industrial Control System (ICS) encompasses the various categories of control systems and integrated components utilized to

control the industrial process. The ICS is facing an increased number of cyber-attacks that had multiple issues. Inefficient safety measures have an adverse impact on the workforce and organization. Some of the effects are production delays, building damage, medical and compensation costs, material losses, loss of business, legal expenses, and tool and equipment damages.

Intrusion detection systems detect the vulnerabilities within network traffic on network infrastructure. It could determine when the hackers begin probing devices that is the initial step to generate secure IIoT [3]. IDS system collects and identifies network traffic, audit data, security logs, and information from system key points to check whether there exist security damages in the network. The solution of IDS for IIoT requires to be customized to the nature of devices. Deep learning methods are used with IoT to improve the efficiency of IIoT applications. It maintains the balance between efficiency and computational cost for next-generation IoT networks.

Various researches provided several techniques for IDSs in industries. This paper provides a survey of multiple existing deep learning techniques for intrusion detection in IIoT. This survey offers various objectives. First, this research describes the preliminary analysis of IIoT systems and Intrusion Detection systems. Then we analyse the different existing deep learning techniques with their advantages and disadvantages for detecting intrusions in IIoT designs. Then we focus on various performance measures and datasets involved in IDS-IIoT. From the integration of surveys, we describe the limitations and challenges of existing methods. Finally, the solutions to rectify these challenges and future directions are sketched. The key contribution of this paper is listed below,

- (a) We review the deep learning-based IDS systems for various industrial IoT applications.
- (b) We present the multiple known cyber datasets, which are utilized in existing researches to classify the intrusions.
- (c) We compare several existing deep learning methods with regard to their advantages and disadvantages.

II. BACKGROUND

A. Intrusion Detection System (IDS)

IDS is used for monitoring the malicious attack in the interconnected network or node. It acts as the line of defence that can protect the node or network from attackers [4]. The malicious activity is defined as the intrusion that is destructive

to sensor nodes. The IDS system can be hardware or software tools. IDS can examine the user actions that determine the known attacks and identifies malicious network behaviour. It analyses the activity of nodes and networks with the determination of various intrusions and alerts the users after detecting the intrusions. So, it's called an alarm or network observer. It eliminates the systems damage by the alert generation before attains the unauthorized attacks. The IDS system can detect both Internal Attacks (IA) and External Attacks (EA). IA's are produced by malicious nodes that interconnected network. EA's are made by third parties that are attained by an external structure. The IDS system observes the network packets and detects whether they are unauthorized users or valid users. IDS involves three stages such as monitoring, investigation and alert. The monitoring component monitors the network patterns, traffics and resources. The study is the core component that determines the intrusions based on the specified algorithm. Alert module raised the alarm when the intrusions are determined. IDSs are categorized into three types which are described below [5],

1) *Host-based IDS System (HIDS)*: HIDS estimates the information determined on a single or multiple host systems which include design, operating system and application files. This system collects the data from internal sources to computer at the operational system level, monitor user activities and monitor executions of system programs. It had the advantages of improved recovery, descriptive logging, fewer false positives, and unknown attack prediction. The disadvantages are unreadable information, complete coverage, indirect information, outsiders, and host interference.

2) *Anomaly-based IDS System (AIDS)*: This system is called as event-based intrusion detection. It determines the malicious behaviours by analysing the event. Initially, it describes the normal activities of the attack. When the activities are varied from normal activity, then it's represented as an intrusion.

3) *Network-based IDS System (NIDS)*: This system identifies the intrusions by monitoring the network traffic through the network interface cards, routers and switches. The data are mainly collected through general network stream like internet packets. NIDS can detect all attacks in LAN and can determine attacks that cannot be achieved by host-based IDS. Advantages of NIDS are ease of deployment, cost, detection range, forensics integrity, and detects all attempts. The disadvantages are a failure at wire speed, direct attack susceptibility, indecipherable packets and Problem of complete coverage .

B. Requirement for IDS in Internet of Things (IoT) Networks

IoT is the growing technology that defines the physical objects that had the capability of exchanging information with other items. These objects communicate with each other without human intervention. IoT is a smart network that communicates all things through the internet for information exchange with approved protocols. Therefore, the user can access anything from anywhere at any time. It utilizes unique addressing methods to communicate these objects and collaborate with objects to generate new services and applications. It presents multiple applications like smart cities, smart homes,

smart environment, health monitoring and smart water [6]. Despite IoT provides various facilities for human routine life based on its reliable and available actions, It requires multi-class security solutions with regard to integrity, privacy and other verification services. The IoT network should be protected against intrusions, and information captured by IoT sensor should be uploaded in an encrypted format. Thus, developing a secure communication is an essential task in the IoT network.

Among the multiple issues of IoT, the security issue cannot be left unnoticed as IoT devices can be retrieved from anywhere through the unauthorized network [7]. When the security problems are not analysed, then the sensitive data may be attacked at any time. So, the security issues are must be identified from the following aspects,

- (a) *Confidentiality*: The attacker can easily interrupt the message passing from source to destination so that the user sensitive information can be leaked and the data can be modified. So, secure information transformation is the most important.
- (b) *Integrity*: The transmit information should be received at the receiver side without any modification. Integrity assures that the information has not been modified by unapproved attackers in transmission.
- (c) *Availability*: Resources should be available when required. The attackers can overflow the resources bandwidth to destruct the accessibility. This accessibility can damage by some malicious attacks.
- (d) *Authenticity*: It performs the identity proof. The users can determine other's identity with that they are in communication. It can be validated through the verification process. So the unauthorized attacks cannot involve in the interaction.
- (e) *Non-repudiation*: It enhances that the sender and receiver are not able to reject the sent and received the information. It provides the proof of origin of the data and integrity of the data.

C. Requirement for IDS in Industrial IoT (IIoT) Networks

IIoT is the revolutionary effort to create smart manufacturing eco-system by using the IoT advantages for industrial process management. IIoT rapidly expands the various industries and services that are discussed in below: In healthcare systems, the IoT devices are used for tracking, sensing, and monitoring of machines, patients and medicines[8]. In the agriculture industry, the IoT devices are utilized for security surveillance of farms, efficient watering of plants, and product storage management. [9].

Transportation and logistics industry plays a vital role in supply chain industries [10]. In this field, the IoT devices are used to determine the vehicle's location for tracking the movements. It's also used to determine the supply time of the product. In the energy sector, the IIoT maintains the supply from and to the grid, billing, and monitoring of leakage. In the mining industry, the IoT devices are used for managing warning systems, sensing disaster signals, tracking underground miner's movement, and monitoring shipments [11]. The strength of the automation industry defines ICS that includes Supervisory Control and Data Acquisition (SCADA)

networks and Programmable Logic Controllers (PLC). Most of the cyber-attacks involve against industrial automated systems such as Stuxnet attack, German steel mill blast furnace attack, Shamoon attack, Mirai, etc..

Numerous cyberattacks are targeted the industrial enterprises around the world. A large amount of vulnerabilities presents in IoT devices for cyber-criminals to attack the industrial process. In traditional, well-protected networks are created with stable defensive mechanisms. Therefore, the robust intrusion detection mechanism is required to fight against attacks and to protect the industrial systems. In the next section, the existing intrusion detections systems are discussed for IIoT by using deep learning.

III. REVIEW ON DEEP LEARNING-BASED IDS APPROACHES

In IIoT, the cyber-attacks are growing at an alarming rate with the increase in connected applications, devices and communication networks. When the attacks occur in IIoT networks, it reduces the availability of systems for end-users, and it increases the theft identity and the number of data breaches, costs and revenue impacts. Though various surveys and researches had been published on IDSs for IoT using deep learning, no survey is present on IDS approaches for IIoT. This section discusses the various deep learning based IDS approaches for IoT and IIoT applications. Some of the discussed deep learning techniques for IDS systems are as follow,

A. Applied for Securing IoT Networks

The authors in [12] have developed an algorithm leveraging the benefits of deep learning to detect DDoS attacks. Also, they have compared the performance of several deep learning approaches over machine learning techniques for DDoS attack detection. The results have confirmed the potential of deep learning to increase the accuracy in detecting DDoS attacks that occur within IoT network. Also, the authors in [13] developed IDS leveraging the potential of convolutional neural networks (CNN) and high performance computing to achieve better intrusion detection performance within IoT networks. In this line, an IDS based on deep belief network based IDS is proposed in [14]. The approach has proved to provide better intrusion detection performance in term of accuracy, F1-score, precision and detection rate on the benchmark dataset, CI-CIDS. Also, the authors in [15] presented deep learning based IDS by stacking nearly five residual networks that pretrained to learn the malicious network behavior and recognize intrusion within IoT network. Several other researchers have attempted to design IDS leveraging the benefits of deep learning from different perspectives as follows

- (a) **Self-taught learning:** The authors in [21] have developed an IDS based on deep learning approaches to recognize four types of attacks that occur within IoT system specially to protect smart-home environment. The approach utilizes self-taught learning framework to analyze six features for attack detection
- (b) **Transfer learning:** The work in [22] adopts feed-forward deep neural networks and transfer learning in encoding the network traffic features to build multi-class and binary

classifiers for recognizing different types of attacks within IoT networks.

- (c) **Ensemble learning:** The work in [23] ensembles set of long term short memory and then employs decision tree to make the final decision in recognizing the attacks within IoT network. The system has proved its effectiveness on real-world datasets with an accuracy of 99% in detection different types of attacks against IoT devices
- (d) **Cloud-based IDS:** The authors in [24] have attempted to incorporate blockchain and deep learning to design a deep blockchain framework for Intrusion detection. Here, authors utilize smart contracts to establish privacy-based blockchain in IoT networks and bidirectional LSTM to analyze network data for intrusion. The framework has proven to serve as a decision support system in supporting the IoT users and cloud providers in securely sharing the data. Similarly, the authors in [25] proposed a IDS based on distributed deep learning approach for detecting different types of attacks within IoT network. The model employs distributed blockchain to detect phishing attacks and LSTM hosted on cloud for detecting botnet attacks within IoT network
- (e) **Fog-Assisted IDS:** The work in [26] proposed an IDS based on recurrent neural networks (RNN) for securing Fog computing from cyberattacks. Here, the system is built using multiple layer of RNN to achieve stability and robustness in protecting the fog computing from cyberattacks. Similarly, the authors in [27] employed DNN to develop network IDS that can be deployed in Fog nodes to detect different types of attacks within Fog assisted IoT network.
- (f) **Botnet IDS:** The authors in [28] utilize convolutional neural networks (CNN) with oversampling and feature engineering techniques to handle effectively the imbalance in intrusion dataset and achieve trade-off in performance between effectiveness and efficiency to detect different types of bot attacks that occur within IoT network. The authors in [29] proposed an IDS integrating the benefits of dendritic cell algorithm and deep learning to select discriminate network traffic features for Botnet attacks. The system has demonstrated its potential in improving the detection rate with reduced false alarm rate for bot-net attacks within IoT networks. The authors in [30] investigates the performance of different deep learning approaches for Botnet attacks within IoT network against machine learning approaches. The deep learning approaches have proved their potential over machine learning approaches for botnet attack detection within IoT networks.

Further, the readers can refer these literature [31], [32], [33] to understand the challenges, solutions and future directions in applying deep learning approaches for IDS within IoT.

B. Applied for Securing Industrial Control System (ICS)

In literature, several machine learning based IDS are proposed for ICS. The state-of-the-art and related works are summarized in Table-I. With the breakthrough evolution of DNN various other fields, the researchers have applied DNN to design IDS for ICS. For example, DNN based IDS is presented for vehicle network security [10]. The system utilizes

TABLE I. SUMMARY ON IDS APPROACHES FOR ICS

Authors	Techniques Adopted	Dataset Used	Detection Performance	Remarks
Liang, Wei, et al [16]	Utilizes multi-feature data clustering and optimization model	NSL-KDD and KDDCup99	Accuracy - 97.8%, and FAR - 8.8%	The model is not evaluated for its efficacy with recent industrial network traffic datasets
Tsang, Chi-Ho et al [17]	Utilizes biologically inspired learning model to extract effective features and enhances clustering based IDS solutions	KDD-Cup99	Accuracy - 92.23%, and FAR - 1.53%	No Efforts are taken to reduce the time complexity of biological inspired learning model. Therein, detection time of the model may be higher than other IDS models.
Jin, Chenglu et al [18]	Applies forward secure logging mechanism for intrusion detection	Proof of concept evaluation	NA	Lightweight IDS model with 54 μ s per scan cycle
Butun, Ismail, et al [19]	Leverages parallel and distributed computing for executing Data streaming applications in intrusion detection	Dataset from real-world AMI	-	Recommends Data streaming paradigm as effective technique for intrusion detection in big industrial networks
Yang, Kai, et al [20]	Deterministic Finite automata uses register status to generate fingerprint of ICS controller and performs both active and passive intrusion detection	Real-worlds experiments are conducted for validation	98% recall rate	The model shows detection rate within 2s

DNN to improve the vehicular network security. The designed DNN is trained with probability-based feature vectors that are retrieved from in-vehicular network packets to learn the network parameters. The system displays the probability of each class as either normal or intrusion based on the malicious traffic packets flows to the vehicle. Also, the system demonstrated higher detection rate for intrusion in the Controller Area Network (CAN) bus. Also, the reference in [34] described the IDS system against malicious attacks on communication network of driverless cars. This research investigated the IDS system for VANET by using ANN that determines the DoS attacks. The focus of this suggested system is to determine the attack through the generated data using network behaviour like trace file. The IDS system utilizes the extracted features as auditable data from the trace file.

Similarly, the reference [9] presented the deep learning method for detecting the intrusions in the agricultural field. This study provided a fast state-of-the-art detector to detect the unknown anomalies. Then the RCNN method is used to attain high accuracy and minimum computation time.

The literature [35] presented a survey on IDS for ICS. They described the various characteristics and updated security needs of ICS. This research defined a new taxonomy for IDS in IDS using multiple techniques such as traffic mining based, protocol analysis-based and control process identification based. They identified the merits and demerits of various classes of IDS and discussed some future developments of IDS system for ICS.

The authors in [36] utilized the Long-short term memory (LSTM) for Omni SCADA intrusion detection. It acts as the data acquisition or supervisor control to detect both temporally

correlated and uncorrelated attacks. The feedforward network (FNN) network determines the temporally uncorrelated attacks with the F1-measures of $99.967 \pm 0.005\%$, and for correlated attacks, it had $58 \pm 2\%$. The combination of FNN and LSTM hybridization method enhanced the IDS performance with $99.68 \pm 0.04\%$ F1 measure. The summary of DNN based IDS for ICS are presented in Table-II.

C. Applied for Securing IIoT Networks

In 2018, two IDS was developed utilizing two different types of deep learning models [41]. The first model utilized deep belief network (DBN). Here, the model was trained and tested with disjoint datasets. In the second model, unlabeled dataset has been utilized to train DBN to learn the changes in intrusion network traffic patterns. In the same year, a secure architecture is introduced.[42] for analyzing SCADA network traffic for intrusion detection and securing ICS equipped with IoT platform. The architecture includes an IDS developed with the ensemble of DBN and SVM. Notably, the architecture utilizes network traffic features and payload features to distinguish normal network traffic from malicious activities. The architecture proved its potential on real SCADA network traffic data with higher detection rate. Also, an IDS was proposed for IIoT utilizing the benefits of deep learning [43]. The system here uses deep feedforward neural network and deep autoencoder for learning the characteristics of malicious network traffic. The system uses information captured from TCP/IP packets to distinguish the intrusion network traffic from normal network traffic behavior. The system was evaluated on NSL-KDD and UNSW-NB15 datasets to demonstrate lower false alarm rate and higher detection rate for intrusion within IIoT system.

TABLE II. SUMMARY ON DEEP LEARNING BASED IDS APPROACHES FOR ICS

Authors	Techniques Adopted	Dataset Used	Detection Performance	Remarks
Li, Beibei, et al [37]	Utilizes federated deep learning scheme based on convolutional neural networks and gated recurrent unit	real industrial CPS	98.64% recall rate	Applicable only for same-domain industrial CPSs
Wang, Zhidong, et al [38]	Utilizes deep neural network with different degrees of discrimination	gas pipeline	100% recall rate all attack types	The model is not evaluated with regard to detection accuracy and detection rate
Leyi, Shi, et al [39]	Utilizes CNN, Bi-LSTM and correlation information entropy	gas pipeline	Accuracy-99.21% and FAR-0.77%	Detection rate of 11.73s
Chu, Ankan et al [40]	Utilizes GoogLeNet to extract features for inception module and LSTM	gas pipeline	Accuracy-97.56% and FAR-2.42%	Adopts attention mechanism for time-series level detection

In 2019 a IDS model leveraging the benefits of unsupervised deep learning methods [44], sparse and noisy deep autoencoder was presented to learn the high level network traffic features and later the network traffic is distinguished using supervised deep learning networks. The proposed model is evaluated for its effectiveness to detect intrusion in IIoT system using dataset collected from remote telemetry streams of gasline system.

In 2020, the authors have proposed an IDS utilizing the advantages of deep random neural network to secure and safeguard IIoT system [1]. But, the system was evaluated on UNSW-NB15 dataset to demonstrate its feasibility and applicability for IIoT. The system displayed higher detection accuracy of 99.54% with low false alarm rate. Similarly, In [45] a fusion based IDS is introduced for securing IIoT. The system partitions the acquired network traffic features into four parts based on the correlation between features. The four group of features namely, content, time-based, host-based and statistics features are transformed to matrix form as an image to enable the processing by four respective CNN for intrusion detection. The system proved its potential for intrusion detection in IIoT system when evaluated with NSL-KDD dataset.

The authors in [2] have attempted to address the significant gap in literature confined to the unavailability of dataset for designing and evaluating IoT/IIoT defense solutions. The authors have presented a representative testbed with seven different sensors and three layers of Cloud, Fog, and Edge to simulate realistic IoT/IIoT system and capture network and intrusion traffic. The new dataset collected from the simulated IoT/IIoT testbed with features to distinguish normal and intrusion network traffic is published under the name TON_IoT and made available for research purpose. Also, the authors have applied several machine learning and deep learning algorithms on the generated TON_IoT dataset as a first-hand evaluation that can serve as a baseline and encourage researchers in this direction. The summary of DNN based IDS for IIoT are presented in Table-III

IV. KEY FINDINGS FROM LITERATURE REVIEW

The aforementioned literature review clearly indicates that plethora of works are published on the application of deep learning approaches for building efficient IDS in IoT environment. Although the industrial sectors have experienced several cyberattack incidents across the world with huge loss, comparatively, less researches have taken place in designing deep learning based IDS for ICS. Also, it can be noticed that very few researches have taken place in IIoT environment when compared to ICS and IoT environment. This clearly indicates that security of IIoT is still in its infancy. Hence, the researchers working in the field of cybersecurity are recommended to focus on the application of deep learning approaches for developing IDS for IIoT environment.

V. LIMITATIONS AND CHALLENGES

The effectiveness of the IDS designed for IIoT system should be evaluated with more experimentations to examine the strong and weak points of detection method in various circumstances. IIoT systems expand a large number of heterogeneous connected devices with power, memory and computational constraints that affect the quality of data. Therefore, some of the challenges of existing deep learning methods in IDS-IIoT are defined as below,

- Deep learning methods demand high computation cost and pose challenge for its implementation on resource-constrained devices to assist onboard security system.
- The deep learning methods are characterized with large number of network parameters and their success depend on training process adopted for network parameter learning. Also it depends on techniques utilized for network parameter initialization.
- IIoT networks are object driven that makes it challenging to apply existing computer networks security mechanisms. Therefore, some specialized tools are required to simulate IIoT environment to capture network traffic and conduct experiment for evaluating the designed ap-

TABLE III. SUMMARY ON DEEP LEARNING BASED IDS APPROACHES FOR IIoT

Authors	Techniques Adopted	Dataset Used	Detection Performance	Remarks
Li, Yanmiao, et al [45]	Utilizes deep learning fusion learning with different single CNN and correlation for feature extraction	NSL-KDD	Accuracy-86.95% and FAR-13.45%	The model is not evaluated for its efficacy with IIoT network traffic datasets
Wu, Di, et al [46]	Utilizes LSTM network and Gaussian Naïve Bayes model	real-life time-series datasets	Accuracy-100%	Model is evaluated for anomaly detection rather than network intrusion
Latif, Shahid, et al [1]	Deep random neural network	UNSW-NB15	Accuracy-99.54%	Model is evaluated for IoT network traffic data but not for industrial IoT network traffic
Muna et al [43]	Utilizes deep autoencoder and deep forwards neural networks	UNSW-NB15 NSL-KDD	Accuracy-99% and FAR-1.8%	Model is evaluated for IoT network traffic data but not for industrial IoT network traffic

proaches with regard to security of IIoT systems from evolving vulnerabilities and threats.

VI. FUTURE DIRECTIONS

In future, the researchers in the field of cybersecurity can focus on the following directions to enhance the security of IIoT,

- Develop IDS model that can enable to enhance the detection performance against unknown attacks
- Lack of sufficient number network traffic samples for training deep learning models for IDS
- Develop lightweight IDS model that can be implemented on resource constrained IIoT devices
- Develop distributed and collaborative IDS model that can safeguard the resources of IIoT from sophisticated attacks
- Develop intrusion detection and prevention system that can secure IIoT from different types of attacks

VII. CONCLUSION

IIoT is the most important part to connect the physical objects to the internet in various industrial applications of the future. During the last decade, the IoT devices usage has rapidly increased due to its capacity of converting objects from application areas into internet hosts. Although, the user's privacy and security are an important challenge due to the security vulnerabilities. Therefore, IoT security must be developed and investigated. The IDS security mechanism is used for IIoT systems and networks with deep learning concept.

In this paper, we presented the various literature survey about deep learning-based IDS for IIoT networks. In this survey paper, we analysed existing papers that were published between 2015 to 2020.

We conclude that research in IDS is still in its incipient and infancy. In addition, it is very hard to develop the comprehensive IDS that can provide more accuracy, robustness, scalability, and protection against all types of intrusions.

ACKNOWLEDGMENT

The authors are very grateful to thank their Deanship of Scientific Research for technical and financial support in publishing this work successfully.

REFERENCES

- Shahid Latif, Zeba Idrees, Zhuo Zou, and Jawad Ahmad. Drann: A deep random neural network model for intrusion detection in industrial iot. In *2020 International Conference on UK-China Emerging Technologies (UCET)*, pages 1–4. IEEE, 2020.
- Abdullah Alsaedi, Nour Moustafa, Zahir Tari, Abdun Mahmood, and Adnan Anwar. Ton_iot telemetry dataset: a new generation dataset of iot and iiot for data-driven intrusion detection systems. *IEEE Access*, 8:165130–165150, 2020.
- Thavavel Vaiyapuri and Adel Binbusayyis. Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: a comparative evaluation. *PeerJ Computer Science*, 6:e327, 2020.
- Adel Binbusayyis and Thavavel Vaiyapuri. Identifying and benchmarking key features for cyber intrusion detection: an ensemble approach. *IEEE Access*, 7:106495–106513, 2019.
- Adel Binbusayyis and Thavavel Vaiyapuri. Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection. *Heliyon*, 6(7):e04262, 2020.
- Thavavel Vaiyapuri. Deep learning enabled autoencoder architecture for collaborative filtering recommendation in iot environment. *CMC-Computers, Materials & Continua*, 68(2):487–503, 2021.
- Adel Binbusayyis and Thavavel Vaiyapuri. Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class svm. *Applied Intelligence*, pages 1–15, 2021.
- Thavavel Vaiyapuri, Adel Binbusayyis, and Vijayakumar Varadarajan. Security, privacy and trust in iomt enabled smart healthcare system: A systematic review of current and future trends. *International Journal of Advanced Computer Science and Applications*, 12(2):731–737, 2021.
- Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft. Deepanomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors*, 16(11):1904, 2016.
- Min-Joo Kang and Je-Won Kang. Intrusion detection system using deep neural network for in-vehicle network security. *PloS one*, 11(6):e0155781, 2016.

- [11] Samyak Jain and K Chandrasekaran. Industrial automation using internet of things. In *Security and Privacy Issues in Sensor Networks and IoT*, pages 28–64. IGI Global, 2020.
- [12] Bambang Susilo and Riri Fitri Sari. Intrusion detection in iot networks using deep learning algorithm. *Information*, 11(5):279, 2020.
- [13] Qasem Abu Al-Haija and Saleh Zein-Sabatto. An efficient deep-learning-based detection and classification system for cyber-attacks in iot communication networks. *Electronics*, 9(12):2152, 2020.
- [14] S Manimurugan, Saad Al-Mutairi, Majed Mohammed Aborokbah, Naveen Chilamkurti, Subramaniam Ganesan, and Rizwan Patan. Effective attack detection in internet of medical things smart environment using a deep belief neural network. *IEEE Access*, 8:77396–77404, 2020.
- [15] Bandar Alotaibi and Munif Alotaibi. A stacked deep learning approach for iot cyberattack detection. *Journal of Sensors*, 2020, 2020.
- [16] Wei Liang, Kuan-Ching Li, Jing Long, Xiaoyan Kui, and Albert Y Zomaya. An industrial network intrusion detection algorithm based on multifeature data clustering optimization model. *IEEE Transactions on Industrial Informatics*, 16(3):2063–2071, 2019.
- [17] Chi-Ho Tsang and Sam Kwong. Multi-agent intrusion detection system in industrial network using ant colony clustering approach and unsupervised feature extraction. In *2005 IEEE international conference on industrial technology*, pages 51–56. IEEE, 2005.
- [18] Chenglu Jin, Saeed Valizadeh, and Marten van Dijk. Snapshotter: Lightweight intrusion detection and prevention system for industrial control systems. In *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*, pages 824–829. IEEE, 2018.
- [19] Ismail Butun, Magnus Almgren, Vincenzo Gulisano, and Marina Papatriantafyllou. Intrusion detection in industrial networks via data streaming. In *Industrial IoT*, pages 213–238. Springer, 2020.
- [20] Kai Yang, Qiang Li, Xiaodong Lin, Xin Chen, and Limin Sun. ifinger: Intrusion detection in industrial control systems via register-based fingerprinting. *IEEE Journal on Selected Areas in Communications*, 38(5):955–967, 2020.
- [21] Marjia Akter, Gowrab Das Dip, Moumita Sharmin Mira, Md Abdul Hamid, and MF Mridha. Construing attacks of internet of things (iot) and a prehensile intrusion detection system for anomaly detection using deep learning approach. In *International Conference on Innovative Computing and Communications*, pages 427–438. Springer, 2020.
- [22] Mengmeng Ge, Naem Firdous Syed, Xiping Fu, Zubair Baig, and Antonio Robles-Kelly. Toward a deep learning-driven intrusion detection approach for internet of things. *Computer Networks*, page 107784, 2021.
- [23] Mahdis Saharkhizan, Amin Azmoodeh, Ali Dehghantanha, Kim-Kwang Raymond Choo, and Reza M Parizi. An ensemble of deep recurrent neural networks for detecting iot cyber attacks using network traffic. *IEEE Internet of Things Journal*, 7(9):8852–8859, 2020.
- [24] Osama Alkadi, Nour Moustafa, Benjamin Turnbull, and Kim-Kwang Raymond Choo. A deep blockchain framework-enabled collaborative intrusion detection for protecting iot and cloud networks. *IEEE Internet of Things Journal*, 2020.
- [25] Gonzalo De La Torre Parra, Paul Rad, Kim-Kwang Raymond Choo, and Nicole Beebe. Detecting internet of things attacks using distributed deep learning. *Journal of Network and Computer Applications*, 163:102662, 2020.
- [26] Muder Almiani, Alia AbuGhazleh, Amer Al-Rahayfeh, Saleh Atiewi, and Abdul Razaque. Deep recurrent neural network for iot intrusion detection system. *Simulation Modelling Practice and Theory*, 101:102031, 2020.
- [27] Nausheen Sahar, Ratnesh Mishra, and Sidra Kalam. Deep learning approach-based network intrusion detection system for fog-assisted iot. In *Proceedings of International Conference on Big Data, Machine Learning and their Applications*, pages 39–50. Springer, 2021.
- [28] Abdelouahid Derhab, Arwa Aldweesh, Ahmed Z Emam, and Farukh Aslam Khan. Intrusion detection system for internet of things based on temporal convolution neural network and efficient feature engineering. *Wireless Communications and Mobile Computing*, 2020, 2020.
- [29] Sahar Ahmed Aldhaheri. Deepdca: Intrusion detection over iot based on artificial immune system and deep learning. 2020.
- [30] S Sriram, R Vinayakumar, Mamoun Alazab, and KP Soman. Network flow based iot botnet attack detection using deep learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 189–194. IEEE, 2020.
- [31] Javed Asharf, Nour Moustafa, Hasnat Khurshid, Essam Debie, Waqas Haider, and Abdul Wahab. A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions. *Electronics*, 9(7):1177, 2020.
- [32] Ankit Thakkar and Ritika Lohiya. A review on machine learning and deep learning perspectives of ids for iot: Recent updates, security issues, and challenges. *Archives of Computational Methods in Engineering*, pages 1–33, 2020.
- [33] Idriss Idrissi, Mostafa Azizi, and Omar Moussaoui. Iot security with deep learning-based intrusion detection systems: A systematic literature review. In *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pages 1–10. IEEE, 2020.
- [34] Khattab M Ali Alheeti, Anna Gruebler, and Klaus D McDonald-Maier. An intrusion detection system against malicious attacks on the communication network of driverless cars. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pages 916–921. IEEE, 2015.
- [35] Yan Hu, An Yang, Hong Li, Yuyan Sun, and Limin Sun. A survey of intrusion detection on industrial control systems. *International Journal of Distributed Sensor Networks*, 14(8):1550147718794615, 2018.
- [36] Jun Gao, Luyun Gan, Fabiola Buschendorf, Liao Zhang, Hua Liu, Peixue Li, Xiaodai Dong, and Tao Lu. Omni scada intrusion detection using deep learning algorithms. *IEEE Internet of Things Journal*, 8(2):951–961, 2020.
- [37] Beibei Li, Yuhao Wu, Jiarui Song, Rongxing Lu, Tao Li, and Liang Zhao. Deepfed: Federated deep learning for intrusion detection in industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics*, 2020.
- [38] Zhidong Wang, Yingxu Lai, Zenghui Liu, and Jing Liu. Explaining the attributes of a deep learning based intrusion detection system for industrial control networks. *Sensors*, 20(14):3817, 2020.
- [39] Shi Leyi, Zhu Hongqiang, Liu Yihao, and Liu Jia. Intrusion detection of industrial control system based on correlation information entropy and cnn-bilstm. *Journal of Computer Research and Development*, 56(11):2330, 2019.
- [40] Ankang Chu, Yingxu Lai, and Jing Liu. Industrial control intrusion detection approach based on multiclassification googlenet-lstm model. *Security and Communication Networks*, 2019, 2019.
- [41] Shamsul Huda, Suruz Miah, John Yearwood, Sultan Alyahya, Hmood Al-Dossari, and Robin Doss. A malicious threat detection model for cloud assisted internet of things (cot) based industrial control system (ics) networks using deep belief network. *Journal of Parallel and Distributed Computing*, 120:23–31, 2018.
- [42] Shamsul Huda, John Yearwood, Mohammad Mehedi Hassan, and Ahmad Almgren. Securing the operations in scada-iot platform based industrial control system using ensemble of deep belief networks. *Applied soft computing*, 71:66–77, 2018.
- [43] AL-Hawawreh Muna, Nour Moustafa, and Elena Sitnikova. Identification of malicious activities in industrial internet of things based on deep learning models. *Journal of information security and applications*, 41:1–11, 2018.
- [44] Muna Al-Hawawreh, Elena Sitnikova, and Frank den Hartog. An efficient intrusion detection model for edge system in brownfield industrial internet of things. In *Proceedings of the 3rd International Conference on Big Data and Internet of Things*, pages 83–87, 2019.
- [45] Yanmiao Li, Yingying Xu, Zhi Liu, Haixia Hou, Yushuo Zheng, Yang Xin, Yuefeng Zhao, and Lizhen Cui. Robust detection for network intrusion of industrial iot based on multi-cnn fusion. *Measurement*, 154:107450, 2020.
- [46] Di Wu, Zhongkai Jiang, Xiaofeng Xie, Xuetao Wei, Weiren Yu, and Renfa Li. Lstm learning with bayesian and gaussian processing for anomaly detection in industrial iot. *IEEE Transactions on Industrial Informatics*, 16(8):5244–5253, 2019.

Sensing and Detection of Traffic Status through V2V Routing Hop Count and Route Energy

Mahmoud Zaki Iskandarani
Faculty of Engineering
Al-Ahliyya Amman University, Amman, Jordan

Abstract—New approach to manage congestion using vehicular communication is presented in this work. The research work using MATLAB simulation, tracked communicating vehicles travelling on roads with constant registration of changes in routes, number of hops, and energy consumed as a function of travelled distances. The area of travel and simulation is divided into blocks or zones to enable sufficient allocation and distribution of Road Side Units (RSUs) that are used to relay communication signals and transmission of Basic Safety Messages (BSMs). The successfully concluded simulation is based on the assumption that as congestion occurs, the number of hops per route and associated energy consumption per transmitted packets will change patterns in terms of hops, routes and consumed energy as traffic passes from low to smooth (optimal) to high density (congestion) states, where at the start of congestion, vehicles start to slow down and become closer to each other in a two dimensional space. The output is used as input to traffic status pattern characterization algorithm (management system) that uses the data to indicate the start of traffic accumulation, thus pre-emptive measures can be taken to avoid congestion and reduction in mobility. The presented analysis proved that it is possible to predict congestion as a function of both hops sequences and consumed energy, depending on the hops pattern which is shown to be symmetric in the case of optimum traffic that flows smoothly. The analysis also showed that when congestion starts to occur, asymmetric hops pattern occurs with hops sequences elements switch and swap places within the identified pattern. Further analysis and polynomial curve fitting proved that congestion control and smooth traffic management using the proposed approach is achievable.

Keywords—V2V; consumed energy; congestion; hops; VANET; routing

I. INTRODUCTION

Mobile data is a critical issue in today's dynamic, fast moving environment. Data transmission requires a network of devices or nodes and links to enable information delivery. One of the most important areas of mobile data application is traffic management and interaction between infrastructure and travelling vehicles. Intelligent transportation system (ITS) enables a more optimized performance of transportation systems through the integration of advanced communication and sensor technologies [1], [2], [3].

A main component of intelligent transportation system is Vehicular Ad Hoc Network (VANET). VANETs are based on similar principles to mobile ad hoc networks (MANETs), where instantaneous formation of a wireless network is carried out in order to enable vehicular data exchange and delivery

within a smart vehicular-infrastructure domain. VANETs are enabled through the use of on-board-units (OBUs) installed in the vehicles and road-side-units (RSUs) installed along side roads [4], [5], [6], [7], [8].

Intelligent transportation systems for vehicular ad-hoc networks motivated a wide range of protective applications, such as vehicle collision warning, lane changing, driver assistance, automatic parking, among others. All of the mentioned applications aim to provide, safety, mobility, efficiency, and comfortable transportation, with main objective of reducing accidents and congestion on the roads that result in many cases from unsafe driving, weather conditions, traffic incidents, or simply from traffic mismanagement. Not to forget drivers conditions, attitudes and experiences [9], [10], [11].

The Road traffic faces congested situations that can lead to chaos. Use of connected vehicles that share information and work cooperatively, would result in safer and greener transportation system. Communication systems used to connect vehicles under cooperative driving principles with intelligent controlling algorithms, plays a critical role in traffic monitoring and management. Such technologies assist in the way that infrastructure is designed and built for a smarter environment. The developed wireless communications standards are now in a state where it can be effectively utilized to enable vehicular communication through Vehicular Ad-Hoc Networking [12], [13].

The main proposed wireless communication technology for vehicular connectivity is the Dedicated Short Range Communication (DSRC), which is mapped through the IEEE 802.11p. Such technology enables both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communication [14], [15], [16], [17]. VANET communication is achieved using dedicated communications devices as part of the On-Board Units (OBUs) in the vehicles, enabling ad hoc networking among vehicles and between vehicles and infrastructure. This will give the ability to monitor vehicles movements and traffic patterns in order to manage safer, less congested and more reliable traffic motion. Effective VANET communications comprises the use of fast and reliable transmission of Basic safety Messages (BSM) to other vehicles on the road. The BSM should be communicated to sufficient number of vehicles in the shortest possible time with considerations to both consumed energy and bandwidth parameters [18], [19], [20], [21].

Wireless ad hoc networks require routing and management protocols, aims at reducing energy consumption together with

self-configuration. They apply real time messaging between vehicles (nodes) using radio wave transmission over specific distance (range). VANETS consist of vehicular mobile nodes, established without a centralized infrastructure, thus, each node will perform the functions of transmitter, receiver, and data router [22], [23], [24], [25], [26], [27], [28].

II. RELATED WORK

Congestion control is a serious issue when dealing with traffic management. Ideally, vehicles would send BSM to other vehicles within a certain range in order to perform certain tasks required by the employed safety application. However, VANETS encounter several challenges due to various obstacles such as transmission delay, available bandwidth, multi-path fading, among others, which affect the performance of a formed network and associated applications [29], [30], [31].

The use of connected and automated vehicles will impose more pressure on reliability, latency, bandwidth conditions, and threshold levels in terms of vehicular communication. Such requirements will be even more stringent when autonomous vehicles are used, especially when entering or leaving roads under different weather and road conditions, during accidents, when changing lanes, among other conditions and under different scenarios [32], [33], [34].

When designing or modifying the management strategies of transportation network, it is essential to evaluate current performance and multiple scenarios, in order to optimize its operations and account for changing capacity in terms of vehicles and also in relation to wireless communication requirements, such as bandwidth, energy, time, and latency. This can be achieved through simulations using various routing algorithms in order to apply optimization to exchanged control message (BSMs) with minimal time delay, and number of hops among others [35], [36].

In this paper, an approach is proposed which uses transmitted bit energy, elapsed signal routing time, and number of hops, to enable better traffic management and congestion control in response to traffic density, which is a function of the increase in the number of vehicles per travelled area over time intervals. The aim here is to use such parameters as input variables to predict and alleviate road congestion, through route change, traffic lights timing change to better manage the existing traffic conditions in relation to infrastructure.

The rest of this paper is divided as follows: Methodology, Results, Analysis and Discussion, Conclusions, References.

III. METHODOLOGY

The main idea of this work is based on the hypothesis that when traffic is very low and traffic density is below a certain threshold with high vehicle mobility (speed, maneuver, lane changing), V2V communication will need to use longer routes with fewer connections between travelling vehicles. This means that the number of hops will increase in response to the reduction in available connections and available routes due to fewer available vehicles (nodes) with abundance of space between the travelling vehicles. However, as the traffic

density grows and number of travelled vehicles increase per travelled distance, it is expected that the number of connections and available communication routes will increase, which result in a decrease in the number of hops per route, as there will be more available connections due to the presence of more vehicles or nodes and less unoccupied spaces on the road. However, when the traffic is optimal and flowing smoothly and efficiently with maximum utilization of the road, then the number of hops and travelled routes should display a unique balanced characteristic that can be used to detect congestion. These assumptions are to be used in this work as indicators to predict congestion and enable better traffic flow and management. This is based on using the following parameters to characterize traffic condition on the roads:

- 1) Number of Vehicles (Nodes): N
- 2) Hops Count: H
- 3) Message Travelled Distance (Route): MTD
- 4) Connections Sequence: CS
- 5) Energy Consumed per Route: ECR

Fig. 1 presents the proposed management system. The system uses the mentioned parameters as inputs to decide if there is a congestion, and act upon such characterization. The simulation traces BSM wireless communication between two specific vehicles (nodes) and analyze the effect of congestion on their connectivity.

Using MATLAB code, the proposed system correlates the number of travelling vehicles (N) to:

- 1) Hops Count (H).
- 2) Message Travelled Distance (MTD).
- 3) Energy Consumed per Route (ECR).
- 4) Connections Sequences (CS).

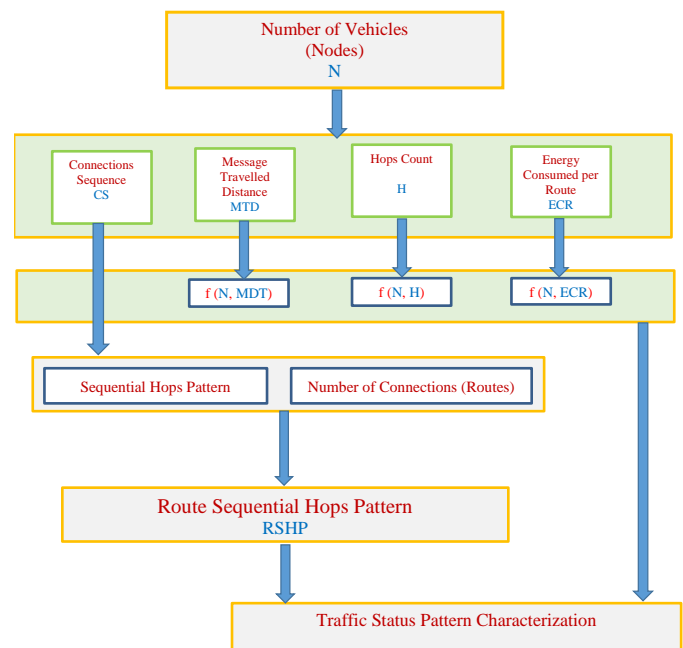


Fig. 1. Proposed Management System.

Connection Sequences are further used to establish connectivity pattern that is implicitly related to Hops Count (H), Message Travelled Distance (MTD), and the Energy Consumed per Route (ECR) through monitoring of repetitive number of hops for each single connection and over the whole travelled route, which is projected onto two parameters:

- 1) Sequential Hops Patterns
- 2) Number of Connections (Routes)

Both Sequential Hops Patterns and Number of Connection (Routes) are then used per number of vehicles in a scenario that is simulated in urban area to produce the Route Sequential Hops Pattern (RSHP) for the state of traffic, enabling traffic density and congestion level characterization through correlation with MTD, H, and ECR all as a function of number of travelling vehicles. This will eliminate errors in traffic characterization and management, when the driver is driving slowly by choice and not as a result of congestion.

The proposed approach eliminates the need to consider travelling time as roads speed differ in different areas and is a function of road design and purpose. Thus making the presented model and simulation applicable to any road type and structure.

As vehicular communication covers both V2V and V2I, with V2V is also accomplished as a result of V2I communication to communicate certain data obtained from Road side Units (RSUs), then it is necessary to establish communication zones over the physical roads to enable determination of transmission ranges within each zone or area of vehicular travel, so that coverage is guaranteed and signal strength is within acceptable limits.

The travelled area or area size is divided into zones of 40 m by 40 m. The proposed and simulated block division allows for addition and operating of RSUs as a function of the total travel distance by vehicles.

The transmission range in one dimension (Longitudinally or laterally) is specified according to equation (1):

$$Transmission\ Range\ (TR) = \beta * ZS \quad (1)$$

Where $\beta \leq 1$ and ZS : Zone Size

The condition of β to be less than one is essential to guarantee non-interference of communicating vehicles groups within each zone and to enable more accurate traffic management especially in the cases of work zones and accidents zones. However, as vehicles are moving there will be no interruption of communications as they move from one zone to the next, very similar to cellular communication that uses cells and handover regions from one coverage to the other. The ratio can be one only if the area is considered as one zone covered by one RSU.

From equation (1), Transmission Range (TR) can be related to Message Travelled Distance (MTD) through equation (2).

$$Transmission\ Range\ (TR) \geq MTD \quad (2)$$

Thus the dynamic MTD for a direct V2V communication between any two vehicles can be approximated as in equation (3).

$$MTD \leq (\beta * ZS) \quad (3)$$

From equation (3), the consumed energy is approximated by equation (4).

$$ECR_{consumed} \leq (ECR * \beta * ZS) \quad (4)$$

Each Hop will consume energy within the specified zone according to equation (5).

$$ECR_{consumed(Hop)} \leq \left(\frac{ECR * \beta * ZS}{H} \right) \quad (5)$$

In this work the used ratio is 0.75 as more than one zone is considered in the simulation as shown in Fig. 2. The transmission range is dynamic and a function of vehicles movements and time. Hence, it is of great advantage to subdivide the travelled area into Zones in case of temporary vehicular groups, where it becomes much easier to manage their communications as they move and leave one group to enter another temporary group that belongs to a different zone.

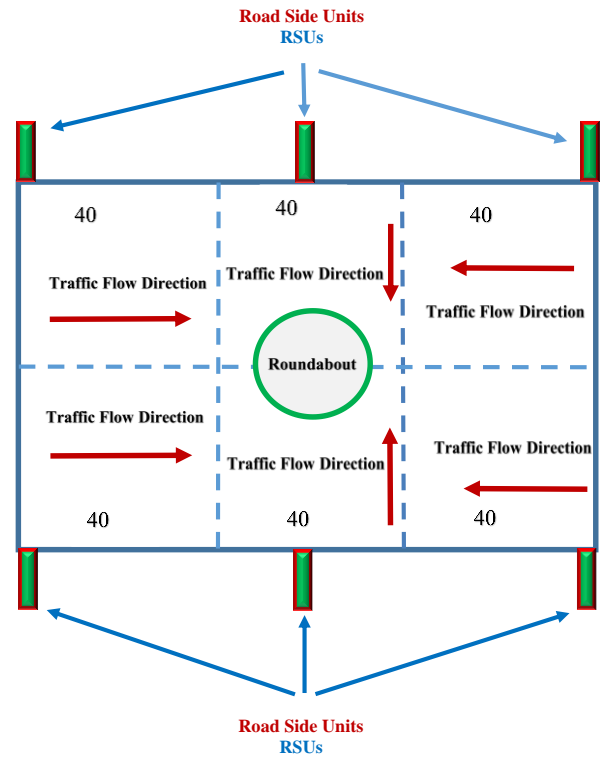


Fig. 2. Simulated Area with Divided Blocks (Zones).

IV. RESULTS

Tables I to III present the initial simulation results with traffic density or vehicles numbers (nodes): {40, 280, 440}.

TABLE I. ENERGY VARIATION AS A FUNCTION OF NUMBER OF CONNECTIONS FOR 40 VEHICLES

Connections Sequence	Hops Count	Message Travelled Distance (m)	Consumed Energy Per Route (J)
1	3	72.83	0.0015
2	3	53.91	0.0009
3	3	52.74	0.0008
4	3	52.29	0.0008
5	3	52.40	0.0008
6	3	52.82	0.0008
7	3	53.43	0.0009
8	3	54.17	0.0009
9	3	55.01	0.0009
10	3	55.95	0.0009
11	3	56.99	0.0010
12	3	58.11	0.0010
13	3	59.32	0.0010
14	3	60.63	0.0011
15	3	62.02	0.0011
16	3	63.50	0.0012
17	3	65.07	0.0012
18	3	66.73	0.0013
19	3	68.46	0.0013
20	3	70.28	0.0014
21	3	72.16	0.0015
22	3	74.12	0.0015
23	3	76.15	0.0016
24	3	78.23	0.0017
25	4	94.34	0.0024
26	4	95.11	0.0024
27	4	96.11	0.0025
28	4	97.31	0.0026
29	4	98.70	0.0026
30	4	90.95	0.0022
31	4	101.57	0.0028
32	4	103.15	0.0029
33	4	104.88	0.0029
34	4	93.74	0.0024

TABLE II. ENERGY VARIATION AS A FUNCTION OF NUMBER OF CONNECTIONS FOR 280 VEHICLES

Connections Sequence	Hops Count	Message Travelled Distance (m)	Consumed Energy Per Route (J)
1	4	110.40	0.0032
2	4	107.98	0.0031
3	4	105.59	0.0030
4	4	103.25	0.0029
5	4	100.97	0.0027
6	4	98.91	0.0026
7	4	97.03	0.0025
8	4	95.17	0.0024
9	4	93.34	0.0024
10	4	94.61	0.0024
11	4	99.45	0.0027
12	4	98.12	0.0026
13	3	87.73	0.0021
14	3	82.18	0.0019

15	3	80.76	0.0018
16	3	79.47	0.0017
17	3	78.31	0.0017
18	3	75.54	0.0016
19	3	80.36	0.0018
20	3	72.04	0.0015
21	3	70.38	0.0014
22	3	69.41	0.0014
23	3	68.06	0.0013
24	3	66.86	0.0013
25	3	65.87	0.0012
26	3	71.58	0.0014
27	3	71.94	0.0015
28	3	76.52	0.0016
29	3	76.47	0.0016
30	2	55.25	0.0009
31	2	53.84	0.0009
32	2	52.50	0.0008
33	2	51.25	0.0008
34	2	50.10	0.0008
35	2	49.07	0.0007
36	2	48.18	0.0007
37	2	47.44	0.0007
38	2	46.86	0.0007
39	2	46.47	0.0007
40	2	46.27	0.0007
41	2	46.27	0.0007
42	2	45.69	0.0007
43	2	45.48	0.0007
44	2	45.31	0.0007
45	2	45.16	0.0007
46	2	45.04	0.0006
47	2	44.94	0.0006
48	2	44.86	0.0006
49	2	44.80	0.0006
50	2	44.76	0.0006
51	2	44.73	0.0006
52	2	44.72	0.0006
53	2	49.27	0.0007
54	2	50.64	0.0008
55	2	52.09	0.0008
56	2	63.41	0.0012
57	2	63.42	0.0012
58	3	60.26	0.0011
59	3	67.24	0.0013
60	3	69.24	0.0014
61	3	81.21	0.0018
62	3	73.24	0.0015
63	4	76.40	0.0016
64	4	77.70	0.0017
65	4	96.36	0.0025
66	4	96.44	0.0025
67	4	96.61	0.0025
68	4	96.86	0.0025
69	4	97.19	0.0025
70	4	97.61	0.0026
71	4	98.13	0.0026

TABLE III. ENERGY VARIATION AS A FUNCTION OF NUMBER OF CONNECTIONS FOR 440 VEHICLES

Connections Sequence	Hops Count	Message Travelled Distance (m)	Consumed Energy Per Route (J)
1	3	67.32	0.0013
2	3	65.98	0.0012
3	3	64.73	0.0012
4	2	63.59	0.0012
5	2	58.77	0.0010
6	2	57.32	0.0010
7	2	55.87	0.0009
8	2	54.49	0.0009
9	2	53.20	0.0009
10	2	52.00	0.0008
11	2	50.91	0.0008
12	2	49.95	0.0008
13	2	49.13	0.0007
14	2	48.47	0.0007
15	2	47.98	0.0007
16	2	47.69	0.0007
17	2	47.59	0.0007
18	2	47.69	0.0007
19	2	47.98	0.0007
20	2	48.47	0.0007
21	2	49.13	0.0007
22	2	49.95	0.0008
23	2	50.91	0.0008
24	2	52.00	0.0008
25	2	53.20	0.0009
26	2	54.49	0.0009
27	2	55.87	0.0009
28	2	57.32	0.0010
29	2	45.39	0.0007
30	2	40.86	0.0006
31	2	41.28	0.0006
32	2	41.78	0.0006
33	2	42.34	0.0006
34	2	42.96	0.0006
35	2	43.66	0.0006
36	2	44.42	0.0006
37	2	45.26	0.0007
38	2	46.18	0.0007
39	2	47.18	0.0007
40	2	48.26	0.0007
41	2	49.43	0.0008
42	2	50.69	0.0008
43	3	52.04	0.0008
44	3	55.22	0.0009
45	3	56.11	0.0009
46	3	57.25	0.0010
47	3	58.59	0.0010
48	3	60.07	0.0011
49	3	61.65	0.0011
50	3	63.31	0.0012
51	3	65.04	0.0012
52	3	66.81	0.0013

53	3	68.61	0.0013
54	3	70.44	0.0014
55	3	72.30	0.0015
56	3	74.18	0.0015
57	3	75.32	0.0016
58	3	76.03	0.0016
59	3	77.91	0.0017
60	3	78.08	0.0017
61	3	79.89	0.0018
62	3	81.86	0.0018
63	3	83.99	0.0019
64	4	86.27	0.0020
65	4	92.70	0.0023
66	4	94.70	0.0024
67	4	96.70	0.0025
68	4	98.24	0.0026
69	4	100.24	0.0027
70	4	102.24	0.0028
71	4	103.00	0.0028
72	4	104.32	0.0029

V. ANALYSIS AND DISCUSSION

Fig. 3 shows the relationship between increasing number of vehicles (nodes) and the average message travelled distance by a transmitted signal from one vehicle to another vehicle. It is clear from the plot that as the traffic density and volume increases per same area size, the travelled distance will decrease, due to vehicles or nodes becoming closer to each other, as their number increases and their speed drops. Hence, the number of 2 hops routes will increase, while 3 hops, and 4 hops routes will decrease, thus reducing the average message travelled distance. This is supported by the detailed hops count and comparative plot shown in Fig. 4.

Fig. 4 present detailed hops distribution that count for each simulated volume of vehicles (nodes) travelling the same areas size. As the plot and associated table present, at 40 vehicles volume, 2 hops routes are not used as the vehicles (nodes) are very far apart (very low traffic), hence only 3 hops and 4 hops connectivity routes are available. As the traffic volume increase to 280 vehicles (nodes), traffic density increases per area size and more vehicles are now able to communicate through 2 hops routes, thus the overall hops count increases in relation to traffic volume with 2 hops routes having larger percentage of the communication routes. However, as the traffic flow is smooth and congestion state is not reached, 3 hops routes and 4 hops routes are almost equally distributed.

With number of vehicles reaching 440 vehicles (nodes) per same area size, the simulated results show a decrease in the 4 hops routes and a large increase in 2 hops routes, with slight increase in 3 hops routes. This is due to the increase in volume and traffic density above normal conditions (optimum level), which lead to vehicles or communicating nodes become closer, slowing down to a limited speed as a result of traffic build up. These observations, which are supported by the carried out MATLAB simulation proves the presented hypothesis and is further backed by the presented plots and associated data in Fig. 5 where, a clear indication of the

reduction in the average consumed energy as the number of hops per route is reduced going from 40 vehicles (nodes) through a smooth (optimum) traffic density per area size (280 vehicles), and up to the start of congestion, where the number of vehicles reached 440 vehicles or nodes.

Such deduction is further proved in Fig. 6, where the corresponding reduction in the average consumed energy per route is shown to correlate with the increase of the 2 hops routes and the reduction of 3 hops routes and 4 hops routes. All plots and analysis are based on data in Tables I to III.

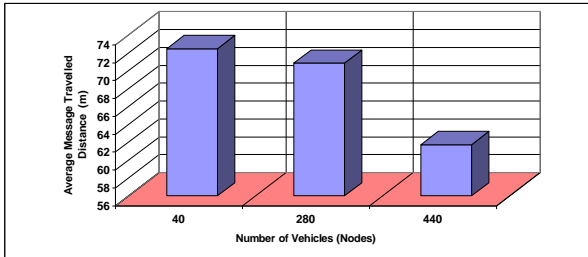


Fig. 3. Effect of Number of Vehicles (nodes) on Average Travelled Distance (Route).

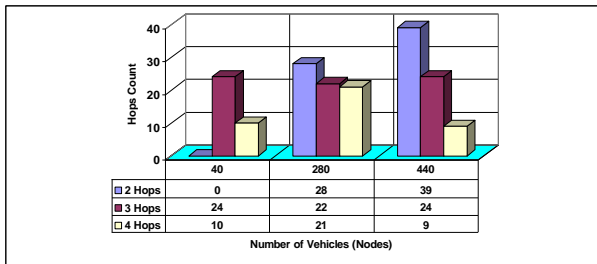


Fig. 4. Relationship between Number of Vehicles (nodes), Hops Distribution and Hop Count.

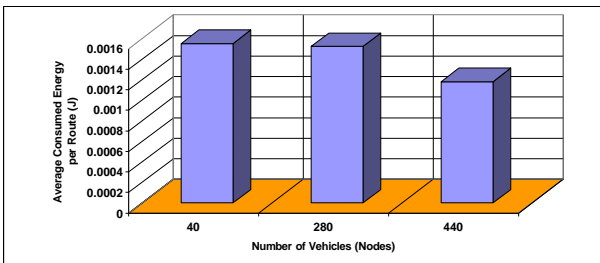


Fig. 5. Effect of Number of Vehicles (nodes) on Average Consumed Energy for Transmitted Bytes.

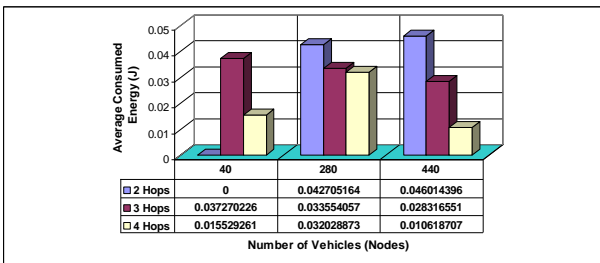


Fig. 6. Relationship between Number of Vehicles (nodes), Hop Distribution and the Average Consumed Energy for Transmitted Bytes.

Based on the previous analysis, the normal pattern for optimum number of vehicles (nodes) in relation to the available area size that results in smooth traffic flow, is related to the sequential hops pattern through symmetrical analysis. A symmetrical transition between routes with different number of hops, indicates smooth traffic with optimum mobility, any deviation from that present the extreme low traffic or very high traffic (congestion). Three cases are depicted by Fig. 7 to 9:

1) If the number of vehicles (nodes) are too low (low traffic flow) in relation to area size: the sequential hops pattern is shown to be [3, 4], (Fig. 7).

2) If the number of vehicles (nodes) are optimum (smooth traffic flow) in relation to area size: the sequential hops pattern is shown to be [4, 3, 2, 3, 4], (Fig. 8).

3) If the number of vehicles (nodes) are high (High density traffic flow (congestion) in relation to area size: the sequential hops pattern is shown to be [3, 2, 3, 4, 3], (Fig. 9).

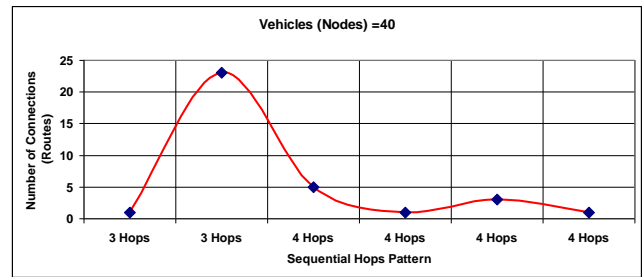


Fig. 7. Hops Distribution Dynamics as a Function of Connections (Routes) for Low Traffic Density.

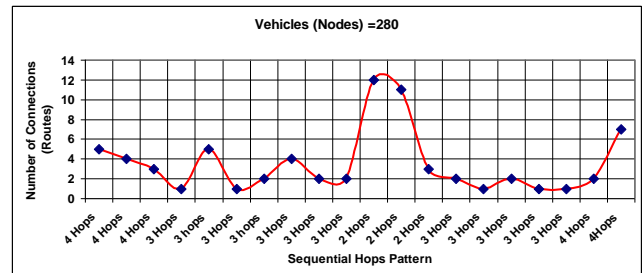


Fig. 8. Hops Distribution Dynamics as a Function of Connections (Routes) for Optimum Traffic Density.

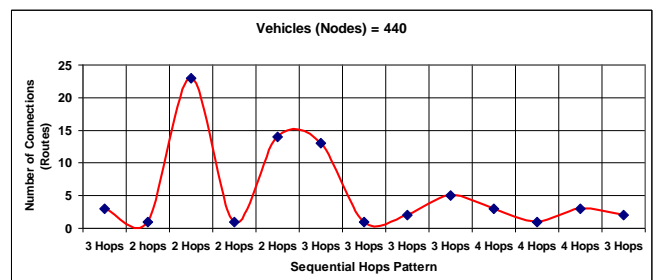


Fig. 9. Hops Distribution Dynamics as a Function of Connections (Routes) for High Traffic Density (Congestion).

Thus, the symmetric case only occurs when the traffic is optimum (case 1), while when it is very low, the 2 hops connectivity does not exist, as the vehicles are far apart (also taking into account speed). In the case of congestion, a shift in the 2 hops is observed as they swap places and a switch also occurs in places between 4 hops and 3 hops routes (case 3). The swapping, switching, and deviation from symmetry implies a serious change in the traffic flow pattern, which is proved through simulation to indicate congestion. The route sequential hops pattern (RSHP), discussed above can be generalized as in equations (6) to (8), assuming maximum allowed number of hops is H_{max} .

$$RSHP_{low\ traffic} = [(H_{max} - 1), \dots, (H_{max})] \quad (6)$$

$$RSHP_{smooth\ traffic} = [(H_{max}), \dots, (H_{max} - 1), (H_{max} - n), (H_{max} - 1), \dots, (H_{max})] \quad (7)$$

$$RSHP_{congestion} = [(H_{max} - 1), (H_{max} - n), (H_{max} - 1), \dots, (H_{max}), \dots, (H_{max} - 1)] \quad (8)$$

$(H_{max} - n)$: The minimum number of hops per route, which is assumed to hold a value of 2.

The expressions in equations (6) to (8) with their associated patterns are correlated to the following variables as indicated in the proposed system in Fig. 1:

- 1) Hops Count (H)
- 2) Message Travelled Distance (MTD)
- 3) Energy Consumed per Route (ECR)

Thus, eliminating any errors that might result when a vehicle is moving at slow speed not because of traffic density, but as a matter of choice.

VI. CONCLUSIONS

This work presented a new concept in congestion control as a function of hops count and route energy variation. The work attempted to solve a critical issue in traffic management and smart cities, namely, congestion. Congestion is a result of inefficient transportation policies, or a consequence of road incidents, will adversely contribute to the economic, environmental, and societal beings. Hence, it is of prime importance to control congestion and work towards smooth traffic flow. As a result of this work, a final correlation can be made between:

- 1) The consumed energy per used route and traffic density.
- 2) The number of connections during vehicle's trip and traffic density.

Fig. 10 and 11 illustrate the relationships with fitted polynomial curves.

The fitted curves mathematical expressions are given in equations (9) and (10).

$$M = 4x10^{-4} N^2 + 273x10^{-3} N + 0.237x10^2 \quad (9)$$

$$ECR = 5x10^{-9} N^2 + 2x10^{-6} N + 15x10^{-4} \quad (10)$$

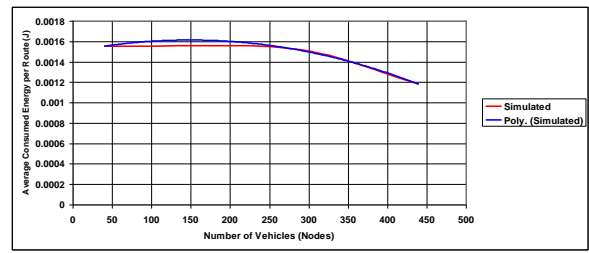


Fig. 10. Effect of Increasing Traffic Density on the Average Consumed Energy Per route.

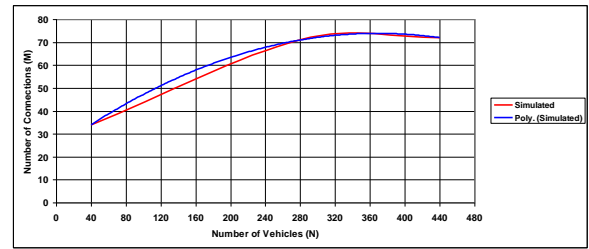


Fig. 11. Effect of Increasing Traffic Density on the Number of Connections Per Trip.

Where;

M: Number of Connections.

ECR: Energy Consumed per Route.

The reduction in energy as nodes number increases is due to slower motion and mobility, which indicates congestion and fewer available connections at the horizon. Also, 2 hops connections are shown to increase as vehicles (nodes) become closer to each other within the available range contributing to lower energy consumption.

Finely, this work achieved the following:

- 1) Establishing a relationship between congestion and number of hops per communication route as a function of travelling vehicles.
- 2) Enabling the correlation between physical travelled distance and communication route length.
- 3) Deducing a relationship between consumed energy and congestion.
- 4) Proving that energy consumed is a function of number of hops, which is also related to the consumed energy that can be correlated to congestion.
- 5) Proposing a correlative algorithm that supports traffic management system.

REFERENCES

- [1] S. Gössling, "ICT and transport behavior: A conceptual review," *Int. J. Sustain. Transp.*, vol. 12, no. 3, pp. 153–164, 2018, doi: 10.1080/15568318.2017.1338318.
- [2] E. Loos, M. Sourbati, and F. Behrendt, "The role of mobility digital ecosystems for age-friendly urban public transport: A narrative literature review," *Int. J. Environ. Res. Public Health*, vol. 17, no. 20, pp. 1–16, 2020, doi: 10.3390/ijerph17207465.
- [3] Z. Pavlović, M. Banjanin, J. Vukmirović, and D. Vukmirović, "Contactless Ict Transaction Model of the Urban Transport Service," *Transport*, vol. 0, no. 0, pp. 1–11, 2020, doi: 10.3846/transport.2020.12529.

- [4] F. Arena and G. Pau, "An overview of vehicular communications," *Futur. Internet*, vol. 11, no. 2, 2019, doi: 10.3390/fi11020027.
- [5] M. El Zorkany, A. Yasser, and A. I. Galal, "Vehicle To Vehicle 'V2V' Communication: Scope, Importance, Challenges, Research Directions and Future," *Open Transp. J.*, vol. 14, no. 1, pp. 86–98, 2020, doi: 10.2174/1874447802014010086.
- [6] Z. El-Rewini, K. Sadatsharan, D. F. Selvaraj, S. J. Plathottam, and P. Ranganathan, "Cybersecurity challenges in vehicular communications," *Veh. Commun.*, vol. 23, p. 100214, 2020, doi: 10.1016/j.vehcom.2019.100214.
- [7] F. Pereira et al., "When Backscatter Communication Meets Vehicular Networks: Boosting Crosswalk Awareness," *IEEE Access*, vol. 8, pp. 34507–34521, 2020, doi: 10.1109/ACCESS.2020.2974214.
- [8] M. S. Sheikh, J. Liang, and W. Wang, "Security and Privacy in Vehicular Ad Hoc Network and Vehicle Cloud Computing: A Survey," *Wirel. Commun. Mob. Comput.*, vol. 2020, 2020, doi: 10.1155/2020/5129620.
- [9] Y. Bai, K. Zheng, Z. Wang, X. Wang, and J. Wang, "MC-Safe: Multi-channel Real-time V2V Communication for Enhancing Driving Safety," *ACM Trans. Cyber-Physical Syst.*, vol. 4, no. 4, 2020, doi: 10.1145/3394961.
- [10] P. Sewalkar and J. Seitz, "Vehicle-to-pedestrian communication for vulnerable road users: Survey, design considerations, and challenges," *Sensors (Switzerland)*, vol. 19, no. 2, 2019, doi: 10.3390/s19020358.
- [11] K. Yu, L. Peng, X. Ding, F. Zhang, and M. Chen, "Prediction of instantaneous driving safety in emergency scenarios based on connected vehicle basic safety messages," *J. Intell. Connect. Veh.*, vol. 2, no. 2, pp. 78–90, 2019, doi: 10.1108/jicv-07-2019-0008.
- [12] S. A. Ahmad, A. Hajisami, H. Krishnan, F. Ahmed-Zaid, and E. Moradi-Pari, "V2V System Congestion Control Validation and Performance," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2102–2110, 2019, doi: 10.1109/TVT.2019.2893042.
- [13] M. Baek, D. Jeong, D. Choi, and S. Lee, "Vehicle trajectory prediction and collision warning via fusion of multisensors and wireless vehicular communications," *Sensors (Switzerland)*, vol. 20, no. 1, 2020, doi: 10.3390/s20010288.
- [14] Venkatamangarao Nampally and Dr. M. Raghavender Sharma, "A Novel Protocol for Safety Messaging and Secure Communication for VANET System : DSRC," *Int. J. Eng. Res.*, vol. V9, no. 01, pp. 391–397, 2020, doi: 10.17577/ijertv9is010029.
- [15] Z. Xu, X. Li, X. Zhao, M. H. Zhang, and Z. Wang, "DSRC versus 4G-LTE for connected vehicle applications: A study on field experiments of vehicular communication performance," *J. Adv. Transp.*, vol. 2017, 2017, doi: 10.1155/2017/2750452.
- [16] Y. A. Vershinin and Y. Zhan, "Vehicle to Vehicle Communication: Dedicated Short Range Communication and Safety Awareness," *2020 Syst. Signals Gener. Process. F. Board Commun.*, pp. 1–3, 2020, doi: 10.1109/IEEECONF48371.2020.9078660.
- [17] S. Kim and B. J. Kim, "Prioritization of Basic Safety Message in DSRC Based on Distance to Danger," *arXiv*, pp. 1–10, 2020.
- [18] B. L. Nguyen, D. T. Ngo, N. H. Tran, M. N. Dao, and H. L. Vu, "Dynamic V2I/V2V Cooperative Scheme for Connectivity and Throughput Enhancement," *IEEE Trans. Intell. Transp. Syst.*, vol. 2020, no. 2, pp. 1–11, 2020, doi: 10.1109/tits.2020.3023708.
- [19] H. Kim and T. Kim, "Vehicle-to-vehicle (V2V) message content plausibility check for platoons through low-power beaconing," *Sensors (Switzerland)*, vol. 19, no. 24, pp. 1–20, 2019, doi: 10.3390/s19245493.
- [20] J. Liu, W. Yang, J. Zhang, and C. Yang, "Detecting false messages in vehicular ad hoc networks based on a traffic flow model," *Int. J. Distrib. Sens. Networks*, vol. 16, no. 2, 2020, doi: 10.1177/1550147720906390.
- [21] X. Liu and A. Jaekel, "Congestion control in V2V safety communication: Problem, analysis, approaches," *Electron.*, vol. 8, no. 5, 2019, doi: 10.3390/electronics8050540.
- [22] C. R. Guerber, E. L. Gomes, M. S. Pereira Fonseca, A. Munaretto, and T. H. Silva, "Transmission Opportunities: A New Approach to Improve Quality in V2V Networks," *Wirel. Commun. Mob. Comput.*, vol. 2019, no. 7, pp. 1–6, 2019, doi: 10.1155/2019/1708437.
- [23] K. Eshteiwi, G. Kaddoum, B. Selim, and F. Gagnon, "Impact of Co-Channel Interference and Vehicles as Obstacles on Full-Duplex V2V Cooperative Wireless Network," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7503–7517, 2020, doi: 10.1109/TVT.2020.2993508.
- [24] H. Chang, Y. E. Song, H. Kim, and H. Jung, "Distributed transmission power control for communication congestion control and awareness enhancement in VANETs," *PLoS One*, vol. 13, no. 9, pp. 1–25, 2018, doi: 10.1371/journal.pone.0203261.
- [25] M. Z. Iskandarani, "Effect of Route Length and Signal Attenuation on Energy Consumption in V2V Communication," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 10, pp. 304–309, 2020, doi: 10.14569/ijacsa.2020.0111039.
- [26] O. S. Eyobu, J. Joo, and D. S. Han, "A broadcast scheme for vehicle-to-pedestrian safety message dissemination," *Int. J. Distrib. Sens. Networks*, vol. 13, no. 11, 2017, doi: 10.1177/1550147717741834.
- [27] S. Li, F. Wang, J. Gaber, and X. Chang, "Throughput and Energy Efficiency of Cooperative ARQ Strategies for VANETs Based on Hybrid Vehicle Communication Mode," *IEEE Access*, vol. 8, pp. 114287–114304, 2020, doi: 10.1109/ACCESS.2020.3003813.
- [28] T. Kim, T. Song, and S. Pack, "An energy efficient message dissemination scheme in platoon-based driving systems," *Energies*, vol. 13, no. 15, pp. 1–23, 2020, doi: 10.3390/en13153940.
- [29] T. Afrin and N. Yodo, "A survey of road traffic congestion measures towards a sustainable and resilient transportation system," *Sustain.*, vol. 12, no. 11, pp. 1–23, 2020, doi: 10.3390/su12114660.
- [30] S. Son and K. J. Park, "BEAT: Beacon inter-reception time ensured adaptive transmission for vehicle-to-vehicle safety communication," *Sensors (Switzerland)*, vol. 19, no. 14, 2019, doi: 10.3390/s19143061.
- [31] D. Punia and R. Kumar, "Experimental Characterization of Routing Protocols in Urban Vehicular Communication," *Transp. Telecommun.*, vol. 20, no. 3, pp. 229–241, 2019, doi: 10.2478/tjt-2019-0019.
- [32] K. Gao, F. Han, P. Dong, N. Xiong, and R. Du, "Connected vehicle as a mobile sensor for real time queue length at signalized intersections," *Sensors (Switzerland)*, vol. 19, no. 9, pp. 1–22, 2019, doi: 10.3390/s19092059.
- [33] Y. Chen, C. Lu, and W. Chu, "A Cooperative Driving Strategy Based on Velocity Prediction for Connected Vehicles with Robust Path-Following Control," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3822–3832, 2020, doi: 10.1109/JIOT.2020.2969209.
- [34] D. F. Xie, Y. Q. Wen, X. M. Zhao, X. G. Li, and Z. He, "Cooperative driving strategies of connected vehicles for stabilizing traffic flow," *Transp. B*, vol. 8, no. 1, pp. 166–181, 2020, doi: 10.1080/21680566.2020.1728590.
- [35] T. Li, D. Ngoduy, F. Hui, and X. Zhao, "A car-following model to assess the impact of V2V messages on traffic dynamics," *Transp. B*, vol. 8, no. 1, pp. 150–165, 2020, doi: 10.1080/21680566.2020.1728591.
- [36] L. Gao, Y. Hou, X. Tao, and M. Zhu, "Energy-Efficient Power Control and Resource Allocation for V2V Communication," *IEEE Wirel. Commun. Netw. Conf. WCNC*, vol. 2020-May, no. 2, pp. 1–2, 2020, doi: 10.1109/WCNC45663.2020.9120612.

Annotated Corpus of Mesopotamian-Iraqi Dialect for Sentiment Analysis in Social Media

AL-KHAFAJI ALI J ASKAR¹, NILAM NUR AMIR SJARIF²

Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia (UTM), Kuala Lumpur, Malaysia

Abstract—Research on Sentiment Analysis in social media by using Mesopotamian-Iraqi Dialect (MID) of Arabic language was rarely found, there is no reliable dataset developed in MID neither an annotated corpus for the sentiment analysis of social media in this dialect. Therefore, this gap was the main stumbling block for researchers of sentiment analysis in MID, for this reason, this paper introduced the development of an annotated corpus of Mesopotamian-Iraqi Dialect for sentiment analysis in social media and named it as (ACMID) stands for (the annotated corpus of Mesopotamian-Iraqi Dialect) to help researchers in future for using this corpus for their studies, to the best of our knowledge this is the first annotated corpus that both classify polarity as well as emotion classification in MID. Likewise, Facebook as the most popular social platform among Iraqis was used to extract the data from its popular Iraqi pages. 5000 comments were extracted from these pages classified by its polarity (Positive, Negative, Neutral, Spam) by two Iraqi annotators, these annotators were simultaneously classifying the same comments according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, Contempt) or no emotion. Cohen's kappa coefficient was then used to compare the two annotators' results to find the reliability of these results. The data shows a comparable value among the two annotators for the polarity classification as high as 0.82, while for the emotion classification the result was 0.65.

Keywords—Sentiment analysis; Mesopotamian dialect; Iraqi dialect; social media; annotated corpus; emotion classification; Arabic language

I. INTRODUCTION

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic among more than 40 million people in Iraq and its neighbors. Making it the second most popular dialect of Arabic after the Egyptian dialect (which reach around 100 million speakers) in the Arab world. Facebook is the most popular social network among Iraqis, and usually, Iraqi people use their dialect in Facebook comments and posts.

Iraq is an important country in the region of the Middle East and the whole world, it is the cradle of civilization and one of the wealthiest countries in the world in its oil reserves and production that might affect the world economy, Iraq was the main front in so many global events during human history, it's hard to find someone in the world does not hear about Iraq because of the events that keep happening there.

Therefore, MID as a dialect for most residents of this country has an important role to extract the opinion of its people to have full knowledge of their thoughts and thinking better than hear their thoughts from others that cannot be

mostly correct and lead to be misleading. Also, understanding people's opinions can be useful in making trading and social decision as well as investing in so many fields of the economy.

Social Media is the main source of getting people's opinions, by extracting data from people's comments and posts useful information can be introduced after classify its polarity and emotion towards certain events and ideas. Facebook as mentioned before is the main platform of social media using by Iraqi people, it has more than 21 million users in Iraq [1], extracting data from Iraqi pages of Facebook can be so useful to get people's thoughts and opinions.

Regardless of the Important of Mesopotamian-Iraqi Dialect (MID) in the world (and Arabic Language in general), studies on Sentiment Analysis in social media using this dialect is so rare and there is no real dataset developed in MID neither an annotated corpus that can be relay on for the sentiment analysis of social media in this dialect [2].

Some Researchers preferred to do their researches on the English version on the original Arabic text instead, because of the complexity of Arabic language in general and the features that facilitates the extracting of the result in the English language to get a more accurate result [3].

Therefore, this gap was the main stumbling block for researchers of sentiment analysis in MID, for this reason, this paper will introduce a new annotated corpus named (ACMID) extracting its data from popular Iraqi Facebook pages to help researchers in the future using this corpus for their studies and researches on sentiment analysis in social media used MID.

To make the new annotated corpus ACMID, Facebook was used to extract the data from its popular Iraqi pages as it is the most popular social platform among Iraqis. 5000 comments were extracted from these pages classified by its polarity (Positive, Negative, Neutral, Spam) by two Iraqi annotators, these annotators were simultaneously classifying the same comments according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, Contempt) or no emotion.

In this paper, related works will be stated in the next section, a brief description for Arabic dialects will be shown in the third section, the fourth section will demonstrate the data collection and pre-processing, the fifth section will state the data annotation and the rules that have to be followed by the annotators, while the sixth section will discuss the results of this work.

II. RELATED WORKS

Related works for sentiment analysis in MID are so limited, most of the related works in the Arabic language are available in MSA and some regional dialects of Egypt (Egyptian dialect), Saudi Arabia (Najidi and Gulf Arabic dialects which referred to as Saudi dialect at most) and other dialects of Arabic language (Levanti, Meghribi, etc.).

AWATEF corpus one of the most reliable corpus by researchers of Arabic, AWATEF corpus was extracting its data from different sources in MSA [4]. COLABA (Cross-Lingual Arabic Blog Alerts) is a project in many Arabic dialects including MID was developing Natural Language Processing (NLP) resources for these dialects [5]. On the other hand, DIWAN software was developed to help training annotators to create their tagging corpus, it can capture the morphological characters in a certain text [6]. Itani et al. build Arabic corpora by extracting their data from Arabic Facebook pages (Al-Arabiyya and the voice) [7].

Al-Kabi et al. [8] create an Arabic corpus from reviews written in MSA and in addition to five Arabic dialects (Egypt dialect, Levant dialect, Arab Peninsula dialect, Maghrebi dialect, and Mesopotamian-Iraqi dialect), this corpus has 250 topics and 1442 reviews.

Meanwhile, many researchers were done studying sentiment analysis in Saudi Arabic dialect, Assiri et al. created the first reliable Saudi annotated corpus from Twitter comments [9]. While SDTC [10] was the first Saudi twitter corpus labeled by three annotators.

Alnawas et al. [11] were one of the few researchers who focuses on MID as the dialect of their interest, they used Doc2Vec to represent for binary classifier of machine learning (Decision Tree, Logistic Regression, Naïve Bayes and Support Vector Machine).

III. MSA, CA/QA AND MID

Modern Arabic Language (MSA) was derived from the Classic Arabic CA in the late 19th century and the beginning of the 20th century by Arab linguistic scholars as a modern form of the CA. MSA is used widely in the Arab world (Arab Homeland as prefer to call by Arabs) as the main language for learning, writing, the conversation among educated people in the universities, legislation, and other formal speech, and sometimes as a lingua franca among Arabs from different dialects of remote regions that cannot be intelligible understood between their speakers (e.g. Iraqi speaking with Algerian).

Classic Arabic Language (CA) or Quranic Arabic (QA) is the root language of all other Arabic dialects. It is based on the text of the Quran (The holy book of Muslims around the world), Quran was first introduced in the 7th century in the west part of the Arabian Peninsula which used the dialect of Arabic of that time in that region as the dialect of Arabic which eventually became the root of all Arabic dialects since.

Most of the Arab speakers cannot distinguish the differences between MSA and CA and most of them consider it as one dialect. Arab people usually named the two dialects as (Al-Arabiyya Al-fusha-العربية الفصحى) [12].

Arabic dialects can be divided into five groups as mention below:

- Mesopotamian Dialects
 - South Mesopotamian Dialect (gelet)
 - North Mesopotamian Dialect (geltu)
- Levantine Dialects
 - North Levantine Arabic
 - Syrian Arabic
 - Lebanese Arabic
 - Çukurova Arabic
 - South Levantine Arabic
 - Jordanian Arabic
 - Palestinian Arabic
- Bedawi Arabic
- Arabian Peninsula Dialects
 - Najdi Arabic
 - Gulf Arabic
 - Bahraini Arabic
 - Hejazi Arabic
 - Yemeni Arabic
 - Omani Arabic
 - Dhofari Arabic
 - Shihhi Arabic
- Egypto-Sudanic Dialects
 - Sudanese Arabic
 - Egyptian Arabic
 - Sa'idi Arabic
 - Chadian Arabic
- Magheribi Dialects
 - Moroccan Arabic
 - Algerian Arabic
 - Tunisian Arabic
 - Libyan Arabic
 - Saharan Arabic
 - Hassaniya Arabic

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic in most of the present-day country of Iraq, some regions in Iraqi neighbors as well as Iraqi people in diaspora around the world. People of this region usually use MID as their mother tongue in their daily conversation while using Modern Standard Arabic MSA in writing, formal conversation, and

media. Using MID in writing was so rare all the time from its development during the last 10 centuries ago until the inventing of the Internet and the phone which was used for texting and chatting at first and then was used when social media came after. South Mesopotamian Dialects (gelet) was used in this work, as it is the main dialect among Iraqis, especially in Baghdad the largest city and the capital of Iraq, Iraqis mostly used this dialect in social media even people from the north part of Iraq [13].

IV. DATA EXTRACTING AND PRE-PROCESSING

Facebook as one of the most popular social media platforms among Iraqi people was used as a source to extract data in Mesopotamian-Iraqi Dialect for sentiment analysis. Three Iraqi Facebook pages were the target to get the data from its comments on different kinds of posts of these pages. The first page called (“دليل مطاعم بغداد”, Baghdad Restaurants Directory (which has more than one million followers, the second page called (“برنامج ولاية بطيخ”, Melon City show) which belongs to a famous comedian show among Iraqis and has more than three million followers, while the third page as an unofficial page of Baghdad university which called (“جامعة بغداد”) and has around forty thousand followers at the time this paper was written.

Facepacer an application for retrieving data from the web was used to extract data from Facebook. At first, getting the address ID of the Facebook page from the Findmyfbid website to specify the page that comments will be retrieved from by Facepacer and then extracting these comments to a CSV file.

In the next step pre-processing of the retrieval data will take place by the following procedures:

- Remove empty comments from the corpus.
- Remove comments that contain just a tagged name without a real review.
- Remove redundancies from the corpus.
- Remove Facebook reactions (like, love, haha, wow, sad, angry).
- Remove serious bad words that cannot be acceptable in any way.
- Remove comments that contain just one character or simple (e.g., “.”, “م”).
- Remove any comment that wasn't written in MID or the Arabic language in general.

V. DATA ANNOTATION

To make the new annotated corpus ACMID two Iraqi Arab native speakers (one doctor in his thirties and one engineer 25 years old) will be involved tagging each comment that was extracted from Facebook pages and classifying them according to their polarity, the polarity classification will be either Positive, Negative or Neutral.

Simultaneously, the annotators will classify these comments according to Ekman's seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise,

Contempt) [14] and if it shows no emotion the annotator will tag it as (no emotion).

The classification of these comments will be done according to the following steps and rules:

- A brief explanation about sentiment analysis will be given to the annotators.
- An example of annotating five comments will be shown to the annotators.
- At first, annotators will be asked to classify ten comments only.
- After that, a short discussion among annotators and their works will take place.
- Annotators will be asked then to complete tagging all the comments separately.
- Annotators will be asked not to discuss their work with each other.
- Annotators will be asked not to influence their personal views about a certain topic in their classification.

VI. RESULTS AND DISCUSSION

The 5000 comments will be classified according to their polarity and emotions by two annotators as mentioned in the previous sections. The polarity will be either positive, negative, neutral or spam, these classifications will give a wide range for the annotators to classify the comments according to their polarity, not limit their choices to the positive or negative classification which might be confusing in some comments for the annotator to choose accordingly.

The second classification is about emotion according to Ekman seven universal emotions (Anger, Fear, Disgust, Happiness, Sadness, Surprise, and Contempt) and if the annotator saw there is no emotion to show in a certain comment, he can then choose the eighth choice which it is (no emotion).

The results of classification according to their polarity for the first annotator shows that positive took 2243 comments out of 5000 with a percentage of 44.86%, while negative took 1682 comments out of 5000 with a percentage of 33.64%, the neutral recorded 1038 out of the 5000 comments with a percentage of 20.76%, and finally the spam recorded only 37 comments out of 5000 comments with a percentage of 0.74%.

The second annotator has the following results, positive recorded 2179 comments out of 5000 with a percentage of 43.58%, negative 1662 comments out of 5000 with a percentage of 33.24%, the neutral recorded 1080 out of the 5000 comments with a percentage of 21.6%, and the spam recorded the same result of the first annotator of 79 comments out of 5000 comments with a percentage of 1.58%.

Table I shows that the annotators agreed on 88.32% for the comment's classification according to their polarity which is considered as so high.

TABLE I. MATRIX ILLUSTRATION FOR THE CONFUSION BETWEEN FIRST AND SECOND ANNOTATORS FOR THE POLARITY CLASSIFICATION

	Positive	Negative	Neutral	Spam	Total
Positive	2034	65	78	2	2179
Negative	60	1501	101	0	1662
Neutral	115	111	850	4	1080
Spam	34	5	9	31	79
Total	2243	1682	1038	37	5000

To ensure the reliability of the result for the polarity classification Cohen Kappa coefficient was used to compare the results between the two annotators, Cohen Kappa is used to measure inter-rater reliability for qualitative items [15], when κ takes into account the possibility of the agreement by chance (AC).

The following formula will show the Cohen Kappa coefficient for the agreement between the two annotators:

$$OA:(2034+1501+850+31)/5000=0.8832$$

$$AC:0.4358*0.4486+0.3324*0.3364+0.216*0.2076+0.0158*0.074$$

$$AC: 0.1955+0.11182+0.04484+0.00012$$

$$AC: 0.35228$$

$$\kappa = (OA-AC) / (1-AC)$$

$$\kappa = (0.8832-0.35228) / (1-0.35228)$$

$$\kappa =0.53092/0.64772$$

$$\kappa =0.8196751682825912$$

The final result for polarity classification shows the Kappa coefficient for the agreement between the two annotators as high as (0.82).

The classification of emotions shows the result for the first annotator as the following: (Anger= “256” out of 5000 comments with a percentage equal to “5.12%”, Fear= “38” out of 5000 comments with a percentage equal to “0.76%”, Disgust= “227” out of 5000 comments with a percentage equal to “4.54%”, Happiness= “976” out of 5000 comments with a percentage equal to “19.52%”, Sadness= “346” out of 5000 comments with a percentage equal to “6.92%”, Surprise=

“336” out of 5000 comments with a percentage equal to “6.72%”, Contempt= “400” out of 5000 comments with a percentage equal to “8%”, and No emotion= “2421” out of 5000 comments with a percentage equal to “48.42%”).

While the result from the second annotator was as the following: (Anger= “369” out of 5000 comments with a percentage equal to “7.38%”, Fear= “45” out of 5000 comments with a percentage equal to “0.9%”, Disgust= “198” out of 5000 comments with a percentage equal to “3.96%”, Happiness= “803” out of 5000 comments with a percentage equal to “16.06%”, Sadness= “360” out of 5000 comments with a percentage equal to “7.2%”, Surprise= “347” out of 5000 comments with a percentage equal to “6.94%”, Contempt= “422” out of 5000 comments with a percentage equal to “8.44%”, and No emotion= “2456” out of 5000 comments with a percentage equal to “49.12%”).

Table II shows that the annotators agreed on 75.06% for the comment’s classification according to their emotions.

Cohen Kappa coefficient again was used to compare the results between the two annotators for the emotion’s classification, the following formula shows the Cohen Kappa coefficient for the agreement between the two annotators:

$$OA:(2004+188+280+168+21+610+243+239)/5000=0.7506$$

$$AC:0.4912*0.4842+0.0738*0.0512+0.0844*0.08+0.0396*0.0454+0.009*0.0076+0.1606*0.1952+0.072*0.0692+0.00694*0.0672$$

$$AC:0.2378+0.0038+0.0068+0.0018+0.0000684+0.0313+0.005+0.0047$$

$$AC: 0.2912$$

$$\kappa = (OA-AC) / (1-AC)$$

$$\kappa = (0.7506-0.2912) / (1-0.2912)$$

$$\kappa =0.4594/0.7088$$

$$\kappa =0.64813769751693$$

The final result for emotion classification shows the Kappa coefficient for the agreement between the two annotators as (0.65).

TABLE II. MATRIX ILLUSTRATION FOR THE CONFUSION BETWEEN FIRST AND SECOND ANNOTATORS FOR THE EMOTION’S CLASSIFICATION

	No-Emotion	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise	Total
No-Emotion	2004	20	27	8	5	327	28	37	2456
Anger	72	188	37	13	3	8	25	23	369
Contempt	63	20	280	24	2	6	11	16	422
Disgust	9	2	12	168	0	2	3	2	198
Fear	8	3	4	0	21	4	4	1	45
Joy	136	6	17	6	3	610	17	8	803
Sadness	67	8	12	7	2	11	243	10	360
Surprise	62	9	11	1	2	8	15	239	347
Total	2421	256	400	227	38	976	346	336	5000

VII. CONCLUSION

Mesopotamian-Iraqi Dialect (MID) is a main dialect of Arabic, Researches that have interested in this dialect were so rare, researchers have difficulties studying sentiment analysis in this dialect because of the lack of reliable annotated corpus in MID as well as a real dataset.

To the best of our knowledge, this paper was introduced the first annotated corpus ACMID that both classify polarity as well as emotion classification in MID. Two annotators were involved to tag the extracted data of comments from three Iraqi famous face pages. The result shows the Kappa coefficient for the agreement between the two annotators for the polarity classification as high as 0.82, while for the emotion classification the result was as 0.65.

Future plan is to applied Machine Learning techniques on the created corpus ACMID (Annotated Corpus of Mesopotamian-Iraqi Dialect).

REFERENCES

- [1] World Population Review, "Facebook Users by Country 2021." <https://worldpopulationreview.com/country-rankings/facebook-users-by-country>.
- [2] M. E. M. Abo, R. G. Raj, and A. Qazi, "A Review on Arabic Sentiment Analysis: State-of-the-Art, Taxonomy and Open Research Challenges," *IEEE Access*, vol. 7, pp. 162008–162024, 2019.
- [3] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster Algorithms on Al-Quran Verses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, 2020, doi: 10.14569/IJACSA.2020.0110832.
- [4] M. Abdul-Mageed and M. T. Diab, "AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis," in *LREC*, 2012, vol. 515, pp. 3907–3914.
- [5] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "COLABA: Arabic dialect annotation and processing," in *Lrec workshop on semitic language processing*, 2010, pp. 66–74.
- [6] F. Al-Shargi and O. Rambow, "Diwan: A dialectal word annotation tool for Arabic," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, 2015, pp. 49–58.
- [7] M. Itani, C. Roast, and S. Al-Khayatt, "Corpora for sentiment analysis of Arabic text in social media," in *2017 8th international conference on information and communication systems (ICICS)*, 2017, pp. 64–69.
- [8] M. Al-Kabi, M. Al-Ayyoub, I. Alsmadi, and H. Wahsheh, "A prototype for a standard arabic sentiment analysis corpus.," *Int. Arab J. Inf. Technol.*, vol. 13, no. 1A, pp. 163–170, 2016.
- [9] A. Assiri, A. Emam, and H. Al-Dossari, "Saudi twitter corpus for sentiment analysis," *Int. J. Comput. Inf. Eng.*, vol. 10, no. 2, pp. 272–275, 2016.
- [10] A. Al-Thubaity, M. Alharbi, S. Alqahtani, and A. Aljandal, "A saudi dialect twitter corpus for sentiment and emotion analysis," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 2018, pp. 1–6.
- [11] A. Alnawas and N. Arici, "Sentiment analysis of iraqi Arabic dialect on Facebook based on distributed representations of documents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–17, 2019.
- [12] L. A. Al Suwaiyan, "Diglossia in the Arabic Language," *Int. J. Lang. Linguist.*, vol. 5, no. 3, pp. 228–238, 2018.
- [13] H. Palva, "From qeltu to galot: Diachronic notes on linguistic adaptation in Muslim Baghdad Arabic," in *Arabic Dialectology*, Brill, 2009, pp. 17–40.
- [14] P. Ekman, "An argument for basic emotions," *Cogn. & Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.
- [15] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochem. medica*, vol. 22, no. 3, pp. 276–282, 2012.

Exploring Parkinson's Disease Predictors based on Basic Intelligence Quotient and Executive Intelligence Quotient

Haewon Byeon

Department of Medical Big Data
College of AI Convergence, Inje University
Gimhae 50834, Gyeongsangnamdo, South Korea

Abstract—It is important to identify the risk factors of dementia and prevent them for the health of patients and caregivers. This study (1) explored sampling methods that could minimize overfitting due to data imbalance using a data-level approach, (2) developed nine ensemble learning models for predicting Parkinson's Disease–Mild Cognitive Impairment (PD-MCI) ((undersampling, oversampling, and SMOTE) × (boosting, bagging, and random forest)=9), and (3) compared the accuracies, sensitivities, and specificities of these models to understand the prediction performance of the developed models. We examined 368 subjects: 320 healthy elderly people (≥ 60 and ≤ 74 years old) without Parkinson's disease (168 men and 152 women) and 48 subjects with PD-MCI (20 men and 28 women). This study used the Cognition Scale for Olde Adults (CSOA), which could measure cognitive functions comprehensively while considering age and education level, to determine the specific cognitive level of the subject. Our study developed nine prediction models ((undersampling, oversampling, and SMOTE) × (boosting, bagging, and random forest)=9) for developing a model to predict PD-MCI based on basic intelligence quotient and executive intelligence quotient. The analysis results showed that a random forest classifier with SMOTE had the best prediction performance with a sensitivity of 69.2%, a specificity of 75.7%, and a mean overall accuracy of 74.0%. In this final model, digit span test-backward, stroop test-interference trial, verbal memory test-delayed recall, verbal fluency test, and confrontation naming test were identified as the key variables with high weight in predicting PD-MCI. The results of this study implied that a random forest classifier with SMOTE could produce models with higher accuracy than a bagging classifier with SMOTE or a boosting classifier with SMOTE when analyzing imbalanced data.

Keywords—Undersampling; oversampling; SMOTE; random forest; Parkinson's disease–mild cognitive impairment

I. INTRODUCTION

The prevalence of dementia is rapidly increasing in South Korea along with the increase of the elderly population [1]. The National Dementia Epidemiology Survey conducted by the Ministry of Health and Welfare in 2012 showed that the dementia prevalence of the elderly (≥ 65 years old) in 2012 was 9.18% and the number of dementia patients was 540,755 (155,955 men and 384,800 women) [2]. The survey predicted that the prevalence of dementia in old age will increase up to 13.17% in 2050 [2]. Dementia is a stressful disease for both patients and their families because the overall cognitive

function of adults who have achieved normal cognitive development declines, the patients have to struggle against dementia for a long time, and symptoms gradually worsen [3]. Therefore, it is important to identify the risk factors of dementia and prevent them for the health of patients and caregivers [4].

Especially, it is critical to screening dementia as soon as possible from the viewpoint of geriatric medicine. Dementia is known as an irreversible disease that is difficult to cure after it occurs [5]. However, thanks to the rapid development of molecular biology, many studies [6,7,8] have continuously reported that cholinergic enzyme inhibitors such as donepezil can delay the progress of dementia or inhibit the decline of cognitive function. As a result, the perception of dementia treatment has been shifted and early detection of high dementia risk groups has emerged as an important topic. Consequently, if we can detect high dementia risk groups sooner, it will be possible to provide professional counseling on the prognosis and help people establish a better health plan in old age.

Before the onset of dementia, the preclinical phase can last from five to seven years [9]. If appropriate therapeutic interventions are provided during this period, it is possible to delay the development of dementia for about 5 years [10]. Therefore, recent studies have focused on detecting the preclinical phase, particularly mild cognitive impairment (MCI), which is known as a middle stage between normal aging and dementia, as soon as possible [11]. Nevertheless, much fewer studies have identified the risk factors of Parkinson's disease–mild cognitive impairment (PD-MCI) [12]. Moreover, it has rarely evaluated the relationship between neuropsychological tests and PD-MCI using machine learning [13].

Over the past decade, many studies have widely utilized ensemble learning, a supervised learning algorithm, for classifying and predicting the complex risk factors of diseases [14,15,16]. Although ensemble learning is known to be more accurate than conventional decision trees [17], when a prediction model is developed using binomial categorical imbalanced data, the recall and precision of it are highly likely to decrease because the classification can be biased into major classes. In particular, in the case of disease data, since the

number of patients is generally smaller than that of healthy people, data imbalanced problems are more likely to occur [18,19]. Therefore, a sampling technique for processing imbalanced data is additionally needed to overcome the prediction error due to class imbalance in disease data. Previous studies [20,21,22] suggested using oversampling, undersampling, and synthetic minority over-sampling technique (SMOTE) to improve the classification performance for imbalanced data. This study (1) explored sampling methods that could minimize overfitting due to data imbalance using a data-level approach, (2) developed nine ensemble learning models for predicting PD-MCI ((undersampling, oversampling, and SMOTE) × (boosting, bagging, and random forest)=9), and (3) compared the accuracies, sensitivities, and specificities of these models to understand the prediction performance of the developed models.

Construction of this study is as follows: Section II explains subjects, measurements, a data-level approach for improving classification performance of imbalanced data, and analyzed variables. Section III compares the results of developed nine prediction model ((undersampling, oversampling, and SMOTE) × (boosting, bagging, and random forest)). Lastly, Section IV presents conclusion and direction for future studies.

II. METHODS

A. Subjects

This study examined 368 subjects: 320 healthy elderly people (≥ 60 and ≤ 74 years old) without Parkinson's disease (168 men and 152 women) and 48 subjects with PD-MCI (20 men and 28 women). In this study, patients with Parkinson's disease were defined as patients diagnosed with idiopathic Parkinson's disease according to the diagnostic criteria of the United Kingdom Parkinson's Disease Society Brain Bank. The criteria for selecting healthy elderly were (1) those who received at least 24 points from the Korean version of Mini-Mental State Examination (K-MMSE)[23], a normal range, (2) those who did not have any impairment in vision and hearing for performing cognitive tests, and (3) those who did not have a history of stroke, Parkinson's disease, or dementia.

G-Power version 3.1.9.6 (Universität Mannheim, Mannheim, Germany) was used to conduct a power test for the sample size of this study. When the number of predictors was 18, significance level (α) was 0.05, power (1-B) was 0.95, and the effect size (f^2) was 0.15, the minimum sample size was estimated as 213. Therefore, the sample size of this study ($n=373$) exceeded the recommended sample size for testing the statistical significance (Fig. 1 & 2).

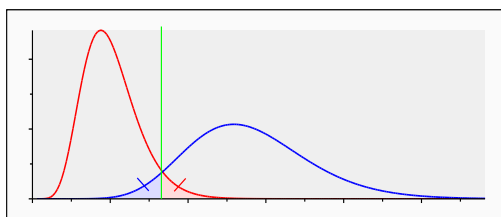


Fig. 1. Results of Power Test.

F Tests - Linear multiple regression: Fixed model, R^2 deviation from zero

Analysis: A priori: Compute required sample size

Input:	Effect size f^2	=	0.15
	α err prob	=	0.05
	Power (1- β err prob)	=	0.95
	Number of predictors	=	18
Output:	Noncentrality parameter δ	=	31.9500000
	Critical F	=	1.6572584
	Numerator df	=	18
	Denominator df	=	194
	Total sample size	=	213
	Actual power	=	0.9508013

Fig. 2. Results of Calculating the Appropriate Sample Size to Test Statistical Significance.

B. Measurements

This study used the Cognition Scale for Olde Adults (CSOA)[24], which could measure cognitive functions comprehensively while considering age and education level, to determine the specific cognitive level of the subject. The CSOA is a standardized cognitive test that can comprehensively measure the cognitive functions of the elderly who are suspected to suffer from cognitive impairment or dementia. The CSOA is composed of stroop simple trial, stroop interference trial, digit span test-forward, digit span test-backward, general information, verbal fluency test, confrontation naming test, Rey complex figure test-copy, recognition, immediately recall, and delayed recall. Among them, stroop simple trial, digit span test-forward, general information, confrontation naming test, and delayed recognition were defined as basic intelligence quotient. Stroop interference trial, digit span test-backward, verbal fluency test, Rey complex figure test-copy, immediately recall, and delayed recall were defined as executive intelligence quotient. The sum of basic intelligence quotient and executive intelligence quotient was defined as full-scale intelligence quotient. Kim (2011)[25] reported that the reliability of the CSOA (Cronbach alpha) was 0.932. This study converted the raw scores of 10 sub-tests into standard scores with an average of 100 and a standard deviation of 15 and used them for machine learning. The composition of the CSOA's sub-tests is presented in Fig. 3.

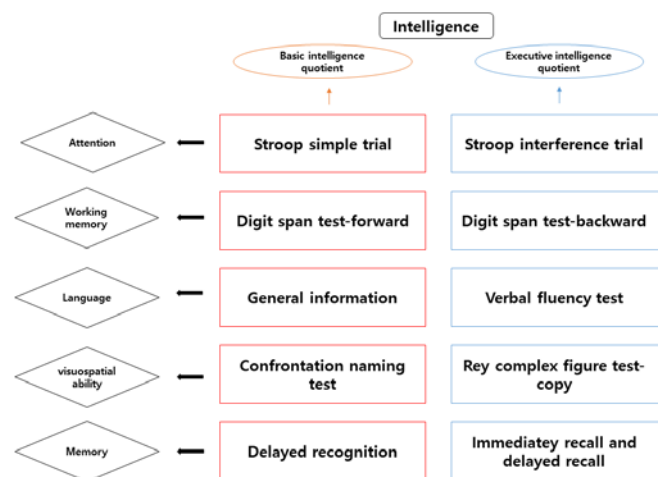


Fig. 3. Composition of basic Intelligence Quotient and Executive Intelligence Quotient in the CSOA.

C. Definitions of Variables

Digit Span Test: When the tester calls out a number, the test subject repeats it immediately after listening to it. There are two-digit span tests: digit span test-forward and digit span test-backward. Each test starts with an item with a short-length number and the length gradually increases as the test progresses. Each raw score is the sum of all items, and it ranges from 0 to 14 points.

Stroop Test: Stroop test consists of a simple trial and an interference trial. The simple trial measures the reaction time takes to tell the colors of 24 circles. The interference trial measures the reaction time to tell the color of a word that describes a color (for example, if “yellow” is written in red, the correct response is “red”). A higher score indicates better reaction sense.

Verbal Memory Test: The verbal memory test is a test that comprehensively calculates the Memory Function Index using 10 picture cards. It is conducted in the order of immediate recall, delayed recall, and recognition. The delayed recall shall be conducted 15-20 after performing the immediate recall. The recognition shall be carried out immediately after completing the delayed recall. The raw score shall be calculated by summing the immediate recall, delayed recall, and recognition, and the score ranges from 0 to 50 points.

General Information: It is a series of questions and answers, and these questions ask common sense. It consists of 20 questions, and each question is one point. Therefore, the total score ranges from 0 to 20 points. A higher score indicates better common sense.

Verbal Fluency Test: It is composed of two trials. The subject shall state nouns in the animal category as many as possible in the first trial and nouns in the crop category as many as possible in the second trial. The time limit for each trial is 1 minute. The raw score is calculated by summing the number of correct responses in the first and second trials. A higher score refers to a better visuospatial function and verbal fluency.

RCFT: Rey Complex Figure (RCF) is to test a subject by copying a figure. Copying is defined as a visuospatial ability, and recalled drawing is defined as visuospatial memory. RCF is evaluated by scoring 18 elements. Each element shall be evaluated by considering the shape and position of each figure, and the original score ranges from 0 to 36 points. A higher score indicates a better visuospatial ability and visuospatial memory.

Confrontation Naming Test: This test asks a subject to read a drawing of an object and say the name (noun) of it. It consists of 24 items, and the range of the raw score is 0 to 24 points. A higher score indicates better confrontation naming ability.

Explanatory variable: Explanatory variables were education level (“middle graduation or below” or “high school graduation or above”), gender (male or female), age, living with a spouse (living together, bereavement/separated, or single), economic activity (yes or no), subjective stress (yes or no), mean monthly household income (<¥1.5 million, ¥1.5-

3.0 million, or \geq ¥3.0 million), smoking (non-smoking or smoking), drinking (non-drinking or drinking), MMSE-K, verbal memory test, stroop test, general information, digit span test, RCFT, confrontation naming test, verbal fluency test, total score of activities of daily living (ADL), and total score of instrumental activities of daily living (IADL).

D. A Data-level Approach for Improving Classification Performance of Imbalanced Data

The results of this study showed that 86.9% (n=320) of the subjects were healthy without suffering from PD-MCI and those suffering from PD-MCI were 13.1% (n=48), indicating that the data was imbalanced. A classifier that learns from binomial categorical imbalanced data, which have a large difference between the size of a major group and that of a minor group, tends to have a classification biased toward the majority group. Therefore, it classifies the majority of the data into the major group to severely reduce the classification accuracy of the minor. In other words, a prediction model developed from unbalanced data can have a higher overall accuracy, but it is highly likely to show a low precision and recall for a minor group. This study used undersampling [26], oversampling [27], and SMOTE methods [28] as data-level approaches to improve the classification performance of binomial categorical imbalanced data.

Undersampling is a method of overcoming the data imbalance issue by randomly removing samples falling in a major class. The undersampling can save time for constructing a model by reducing the amount of data, but it has a disadvantage of losing information [20,29]. Oversampling is a method of overcoming the data imbalance issue by randomly replicating samples falling in a minor class [30].

The oversampling technique takes more time to build a model because the sample size increases, and it may cause an overfitting issue because it copies samples in a minor class [22,31]. The SMOTE finds n nearest neighbors in a minor class of a certain datum in the minor class. Afterward, it draws a line between the datum and the nearest neighbor and randomly generates data along the line until these randomly generated data become synthetic [32].

E. Development of Prediction Models and Evaluation of Prediction Performance

This study developed nine prediction models ((undersampling, oversampling, and SMOTE) \times (boosting, bagging, and random forest)=9) for developing a model to predict PD-MCI based on basic intelligence quotient and executive intelligence quotient. The prediction performance of the developed models was tested by using 5-fold cross-validation. Since the ensemble algorithm has randomness, when the ensemble model was reiterated, seed #12468 was always used. In all ensemble models, the number of decision trees (ntree) was set to 100.

The prediction performance of the developed models was compared by using the accuracy, sensitivity, and specificity of each model. Accuracy indicates the rate of predicting the outcome correctly. Sensitivity refers to the rate of predicting PD-MCI as PD-MCI. Specificity means the rate of predicting a healthy elderly person without PD-MCI and a healthy

elderly person without PD-MCI. This study compared the prediction performance of models and defined that the best prediction model was a model with the highest accuracy while sensitivity and specificity were at least 0.6. The best model was selected as the final model for predicting PD-MCI. All analyzes were performed using R version 4.0.2 (Foundation for Statistical Computing, Vienna, Austria).

III. RESULTS

A. Comparing the Accuracy of the Developed Prediction Models

The accuracy, sensitivity, and specificity of the nine prediction models are presented in Fig. 4, 5, and 6, respectively. The analysis results showed that a random forest classifier with SMOTE had the best prediction performance with a sensitivity of 69.2%, a specificity of 75.7%, and a mean overall accuracy of 74.0%. On the other hand, a boosting classifier with undersampling had the worst performance among the nine prediction models with a sensitivity of 51.8%.

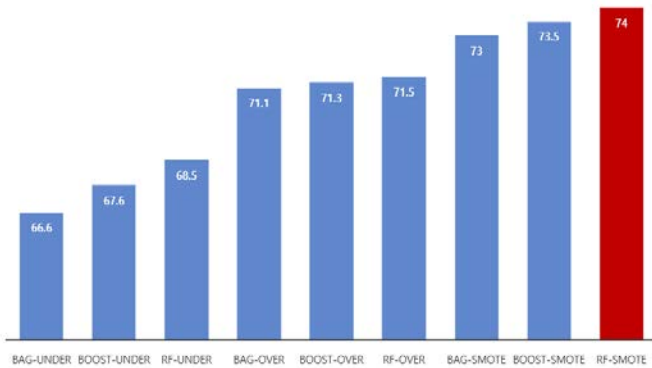


Fig. 4. Accuracy of the Nine Classifier Ensembles.

RF-SMOTE=random forest classifier with SMOTE; RF-OVER=random forest classifier with oversampling; RF-UNDER=random forest classifier with undersampling; BAG-SMOTE=bagging classifier with SMOTE; BAG-OVER=bagging classifier with oversampling; BAG-UNDER=bagging classifier with undersampling; BOOST-SMOTE=boosting classifier with SMOTE; BOOST-OVER=boosting classifier with oversampling; BOOST-UNDER=boosting classifier with undersampling.



Fig. 5. Sensitivity of the Nine Classifier Ensembles.

RF-SMOTE=random forest classifier with SMOTE; RF-OVER=random forest classifier with oversampling; RF-UNDER=random forest classifier with undersampling; BAG-SMOTE=bagging classifier with SMOTE; BAG-OVER=bagging classifier with oversampling; BAG-UNDER=bagging classifier with undersampling; BOOST-SMOTE=boosting classifier with SMOTE; BOOST-OVER=boosting classifier with oversampling; BOOST-UNDER=boosting classifier with undersampling.

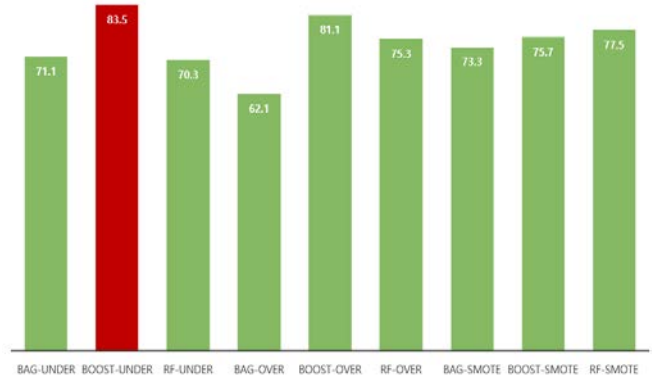


Fig. 6. Specificity of the Nine Classifier Ensembles.

RF-SMOTE=random forest classifier with SMOTE; RF-OVER=random forest classifier with oversampling; RF-UNDER=random forest classifier with undersampling; BAG-SMOTE=bagging classifier with SMOTE; BAG-OVER=bagging classifier with oversampling; BAG-UNDER=bagging classifier with undersampling; BOOST-SMOTE=boosting classifier with SMOTE; BOOST-OVER=boosting classifier with oversampling; BOOST-UNDER=boosting classifier with undersampling.

B. Importance of Variables for PD-MCI Classification in the Final Model

The normalized importance of the variables of the final model (random forest classifier with SMOTE) is presented in Fig. 7. In this model, digit span test-backward, stroop test-interference trial, verbal memory test-delayed recall, verbal fluency test, and confrontation naming test were identified as the key variables with high weight in predicting PD-MCI. Among them, digit span test-backward was the most important variable in a random forest classifier with SMOTE.

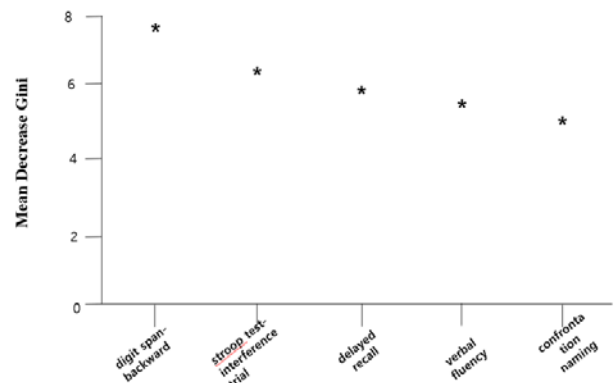


Fig. 7. The Importance of Variables in the Random Forest Classifier with SMOTE-based PD-MCI Prediction Model (Only the Top 5 are Presented).

IV. DISCUSSION

This study compared the prediction performance of nine ensemble learning models ((undersampling, oversampling, and SMOTE) × (undersampling-boosting, bagging, and random forest)=9) for predicting PD-MCI. The results of this study showed that the random forest classifier with SMOTE was the best model (sensitive=69.2%, specificity=75.7%, and mean overall accuracy=74.0%). The result of this study agreed with the results of previous studies [16, 33] showing that random forest based models were superior to other machine learning algorithms for predicting diseases. Particularly, this study developed models by applying oversampling, undersampling, and SMOTE as data-level approaches for improving the classification performance of imbalanced data. It is noteworthy that the accuracy of a random forest classifier with SMOTE was better than that of other learning machine algorithms and ensemble models with SMOTE, oversampling-random forest, or undersampling-random forest.

V. CONCLUSION

The results of this study implied that a random forest classifier with SMOTE could produce models with higher accuracy than a bagging classifier with SMOTE or a boosting classifier with SMOTE when analyzing imbalanced data. Additional studies are needed to compare the accuracy by using various datasets from diverse fields to prove the prediction performance of a random forest classifier with SMOTE.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091, NRF-2019S1A5A8034211).

REFERENCES

- [1] M. J. Cho, The prevalence and risk factors of dementia in the Korean elderly. Health and welfare policy forum, vol. 2009, no. 10, pp. 43-48, 2009.
- [2] G. J. Yoon, The 2021 outlook for healthcare policy. Health and welfare policy forum, vol. 2021, no. 1, pp. 9-22, 2021.
- [3] J. Hashimie, S. K. Schultz, and J. T. Stewart, Palliative care for dementia: 2020 update. Clinics in Geriatric Medicine, vol. 36, no. 2, pp. 329-339, 2020.
- [4] H. Byeon, Development of depression prediction models for caregivers of patients with dementia using decision tree learning algorithm. International Journal of Gerontology, vol. 13, pp. 314-319, 2019.
- [5] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & Dementia, vol. 9, no. 1, pp. 63-75, 2013.
- [6] P. Celsis, Age-related cognitive decline, mild cognitive impairment or preclinical Alzheimer's disease?. Annals of Medicine, vol. 32, pp. 6-14, 2000.
- [7] S. T. DeKosky, and K. Marek, Looking backward to move forward: early detection of neurodegenerative disorders. Science, vol. 302, no. 5646, pp. 830-834, 2003.
- [8] A. Lleó, S. M. Greenberg, and J. H. Growdon, Current pharmacotherapy for Alzheimer's disease. Annual Review of Medicine, vol. 57, pp. 513-533, 2006.
- [9] K. M. Langa, and D. A. Levine, The diagnosis and management of mild cognitive impairment: a clinical review. JAMA, vol. 312, no. 23, pp. 2551-2561, 2014.
- [10] T. Luck, M. Luppia, S. Briel, and S. G. Riedel-Heller, Incidence of mild cognitive impairment: a systematic review. Dementia and Geriatric Cognitive Disorders, vol. 29, no. 2, pp. 164-175, 2010.
- [11] J. M. Ellison, A 60-year-old woman with mild memory impairment: review of mild cognitive impairment. JAMA, vol. 300, no. 13, pp. 1566-1574, 2008.
- [12] G. J. Geurtsen, J. Hoogland, J. G. Goldman, B. A. Schmand, A. I. Tröster, D. J. Burn, and I. Litvan, Parkinson's disease mild cognitive impairment: application and validation of the criteria. Journal of Parkinson's Disease, vol. 4, no. 2, pp. 131-137, 2014.
- [13] H. Byeon, Predicting the severity of Parkinson's disease dementia by assessing the neuropsychiatric symptoms with an SVM regression model. International Journal of Environmental Research and Public Health, vol. 18, no. 5, pp. 2551, 2021.
- [14] K. H. Miao, J. H. Miao, and G. Miao, Diagnosing coronary heart disease using ensemble machine learning. International Journal of Advanced Computer Science and Application, vol. 7, no. 10, pp. 30-39, 2016.
- [15] A. K. Verma, S. Pal, and S. Kumar, Classification of skin disease using ensemble data mining techniques. Asian Pacific Journal of Cancer Prevention, vol. 20, no. 6, pp. 1887-1894, 2019.
- [16] H. Byeon, Best early-onset Parkinson dementia predictor using ensemble learning among Parkinson's symptoms, rapid eye movement sleep disorder, and neuropsychological profile. World Journal of Psychiatry, vol. 10, no. 11, 245-259, 2020.
- [17] H. Byeon, Evaluating the accuracy of models for predicting the speech acceptability for children with cochlear implants. International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 25-29, 2021.
- [18] H. He, and E. A. Garcia, Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.
- [19] H. Byeon, Comparing the accuracy and developed models for predicting the confrontation naming of the elderly in South Korea using weighted random forest, random forest, and support vector regression. International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 326-331, 2021.
- [20] H. Byeon, Development of a physical impairment prediction model for Korean elderly people using synthetic minority over-sampling technique and XGBoost. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 36-41, 2021.
- [21] R. Mohammadi, R. Javidan, M. Keshtgari, and N. Rikhtegar, SMOTE: an intelligent SDN-based multi-objective traffic engineering technique for telesurgery. IETE Journal of Research, vol. 1-11, 2021.
- [22] H. Byeon, Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. International Journal of Advanced Computer Science and Applications, vol. 12, no. 1, pp. 74-79, 2021.
- [23] Y. C. Kwon, Korean version of mini-mental state examination (MMSE-K). Journal of the Korean Neurological Association, vol. 1, pp. 123-135, 1989.
- [24] H. K. Kim, and T. Y. Kim, Cognition scale for older adults; CSOA: manual. Neuropsy Incorporated, Daegu, 2007.
- [25] Y. S. Kim, Diabetes and cognitive function in community-dwelling older adults. Journal of Korean Academy of Community Health Nursing, vol. 22, no. 4, pp. 377-388, 2011.
- [26] S. J. Yen, and Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions. Expert Systems with Applications, vol. 36, no. 3, pp. 5718-5727, 2009.
- [27] L. Abdi, and S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques. IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 1, 238-251, 2015.
- [28] 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.
- [29] H. Guan, Y. Zhang, M. Xian, H. D. Cheng, and X. Tang, SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. Applied Intelligence, vol. 51, no. 3, pp. 1394-1409, 2021.

- [30] M. Wang, X. Yao, and Y. Chen, An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients. *IEEE Access*, vol. 9, pp. 25394-25404, 2021.
- [31] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, Diabetes detection using deep learning techniques with oversampling and feature augmentation. *Computer Methods and Programs in Biomedicine*, vol. 202, pp. 105968, 2021.
- [32] B. Chen, S. Xia, Z. Chen, B. Wang, and G. Wang, RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise. *Information Sciences*, vol. 553, pp. 397-428, 2021.
- [33] A. Sarica, A. Cerasa, and A. Quattrone, Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Frontiers in Aging Neuroscience*, vol. 9, pp. 329, 2017.

Is Deep Learning Better than Machine Learning to Predict Benign Laryngeal Disorders?

Haewon Byeon

Department of Medical Big Data
College of AI Convergence, Inje University
Gimhae 50834, Gyeongsangnamdo, South Korea

Abstract—It is important in otolaryngology to accurately understand the etiology of a laryngeal disorder, diagnose it early, and provide appropriate treatment accordingly. The objectives of this study were to develop models for predicting benign laryngeal mucosal disorders based on deep learning, naive Bayes model, generalized linear model, a Classification and Regression Tree (CART), and random forest using laryngeal mucosal disorder data obtained from a national survey and confirm the best classifier for predicting benign laryngeal mucosal disorders by comparing the prediction performance and runtime of the developed models. This study analyzed 626 subjects (313 people with a laryngeal disorder and 313 people without a laryngeal disorder). In this study, deep learning was the best model with the highest accuracy (0.84). However, the runtime of deep learning was 39min 41sec, which was a 10 times longer development time than CART (3min 7sec). This model confirmed that subjective voice problem recognition, pain and discomfort in the last two weeks, education level, occupation, mean monthly household income, high-risk drinker, and current smoker were major variables with high weight for the benign laryngeal mucosal disorders of Korean adults. Among them, subjective voice problem recognition was the most important factor with the highest weight. The results of this study implied that the prediction performance of deep learning could be better than that of machine learning for structured data, such as health behavior and demographic factors as well as video and image data.

Keywords—Benign laryngeal mucosal disorder; voice disorder; deep learning; Naive Bayes model; generalized linear model

I. INTRODUCTION

Laryngeal disorders include organic dysphonia, caused by the structural changes (anatomical changes) of the larynx including the vocal cords, and functional dysphonia, which changes voice due to health risk behaviors (e.g., smoking or drinking) and improper habits (e.g., abuse or misuse of voice) [1]. In particular, benign laryngeal disorders refer to laryngeal disorders except for laryngeal cancer, a malignant tumor [2]. They are caused by abnormalities in the nervous system, mucous membranes, and cartilage [3], and they are frequently found in the adult population [4]. Benign laryngeal disorders include vocal polyp, vocal nodule, vocal cyst, Reinke's edema, vocal sulcus, vocal scar, contact granuloma, and laryngeal papilloma [5][6].

The prevalence of laryngeal disorders was 6.6% based on the American population [7]. Roy et al. (2005) reported that at least 1 in 10 Americans had experienced voice problems at

least once in their lifetime [7]. There is not enough data regarding the prevalence of laryngeal disorders in South Korea. The Otolaryngology Examination Survey of the 2012 Korean National Health and Nutrition Survey reported that the prevalence of benign laryngeal disorders was approximately 2.5% in South Korea [8]. It was reported that the prevalence of laryngeal disorders is higher among men than women and smokers than nonsmokers [9][10]. It was also reported that the risk of laryngeal disorders was 1.4 to 1.6 times higher in managers, professionals, and service & sales workers than economically inactive people [11][12].

Voice is a very critical function for maintaining daily life. Particularly, it is directly related to living for certain occupations such as teachers, announcers, and singers. Consequently, discovering a laryngeal disorder early for maintaining a healthy voice can greatly improve the quality of patients' life [13,14,15]. Therefore, it is important in otolaryngology to accurately understand the etiology of a laryngeal disorder, diagnose it early, and provide appropriate treatment accordingly.

To date, the most common risk factors causing benign laryngeal mucosal disorders are voice abuse and wrong vocalization habit [16,17,18,19,20,21,22]. Other very diverse factors (e.g., smoking, drinking, viral infection, upper respiratory tract infection, and laryngopharyngeal reflux) have also been reported as risk factors [16,17,18,19,20,21,22]. However, since a disease is a result of complex interactions between multiple risk factors, not caused by a single risk factor, it is limited to predict a disease by exploring only individual risk factors [23]. To make it harder, different treatments need to be given according to individual characteristics (habits) and etiology, even though the shape of the lesions of a laryngeal disorder on the vocal cord mucosa is similar [24]. Consequently, it is important to fully understand the etiology of a benign laryngeal mucosal disorder and identify multiple risk factors of the disease in order to perform accurate diagnosis and treatment. Nevertheless, most studies that have evaluated the risk factors of laryngeal disorders have just tried to find individual risk factors using regression analysis [25,26,27,28,29], and only a few studies have explored the multiple risk factors of benign laryngeal mucosal disorders using machine learning [30].

Supervised learning-based machine learning has been used as a way to detect a disease and identify multiple risks in recent years [31,32,33]. Many recent studies [34,35] have reported that neural network-based deep learning is more accurate in

classifying and predicting diseases than machine learning. Nevertheless, previous studies [36, 37] mainly focused on developing classifiers for discriminating the presence of laryngeal diseases by mostly using video and image data. However, there are not enough studies on developing models to predict benign laryngeal mucosal disorders while reflecting various features (e.g., health behavior, disease, and demographic characteristics) in health surveys. The objectives of this study were to develop models for predicting benign laryngeal mucosal disorders based on deep learning, naive Bayes model, generalized linear model, a Classification and Regression Tree (CART), and random forest using laryngeal mucosal disorder data obtained from a national survey and confirm the best classifier for predicting benign laryngeal mucosal disorders by comparing the prediction performance and runtime of the developed models.

Construction of this study is as follows: Section II explains data source, measurements, development and validation of prediction models. Section III compares the results of developed machine learning models. Lastly, Section IV presents conclusion and direction for future studies.

II. MATERIALS AND METHODS

A. Data Source

This study targeted adults (≥ 19 years old) who participated in the otolaryngology examination and completed the 2012 KNHANES. The KNHANES extracts survey plots using the proportional allocation systematic sampling method that stratifies administrative districts and types of residences across the country and extracts samples proportional to the population survey plots of each layer. This study selected 4,528 adults (313 subjects with a laryngeal disorder and 4,215 subjects without a laryngeal disorder) who completed the health questionnaire, the otolaryngology questionnaire, and laryngeal endoscopy as the primary subjects of this study. Since the prevalence of a laryngeal disorder was only 6.9% among the subjects, showing a data imbalance issue, this study resolved the imbalance issue by using propensity score matching, which matched sex and age (1:1 ratio). Finally, this study analyzed 626 subjects (313 people with a laryngeal disorder and 313 people without a laryngeal disorder).

This study conducted a power test using G-Power program 3.1.9 (Universität Mannheim, Mannheim, Germany) for the final analysis data. When power (1-B) was 0.95, significance level (alpha) was 0.05, effect size (f2) was 0.35, and 201 predictor variables were applied, the appropriate sample size was 361. Therefore, the sample size of this study (626) satisfied the appropriate sample size for testing statistical significance (Fig. 1).

B. Variables

Benign laryngeal disease [20] in this study were defined as vocal nodules, laryngeal polyps, intracordal cysts, reinke's edema, laryngeal granuloma, glottic sulcus and laryngeal keratosis (Fig. 2). The explanatory variables were occupation (economically-inactive, non-manual, manual), educational level (elementary school graduates and lower, junior high school graduates, high school graduates, college graduates and over), high-risk drinking (yes, no),

Income (quartile), smoking (current smoker, previous smoker, or non-smoker), skipped yesterday's breakfast (yes or no), skipped yesterday's lunch (yes or no), skipped yesterday's dinner (yes or no), dietary supplement consumption in the past one year (yes or no), usual fluid intake (g), protein intake (g), fat intake (g), carbohydrate intake (g), calculus intake (g), sodium intake (g), sinusitis prevalence (yes or no), otitis media prevalence (yes or no), tinnitus prevalence (yes or no), depression for two consecutive weeks (yes or no), pain and discomfort in the last two weeks (yes or no), and subjective voice problem recognition (yes or no).

C. Development and Validation of Prediction Models

This study developed models for predicting benign laryngeal disorders using deep learning, naive Bayes model, generalized linear model, CART, and random forest and compared the accuracy and runtime of them to check their prediction performance. Since this study had a small sample size ($n=626$), it could deteriorate the reliability when evaluating the prediction performance using held-out validation. Therefore, this study carried out 5-fold cross-validation to evaluate the prediction performance (Fig. 3). The R code of the 5-fold cross-validation is shown in Fig. 4.

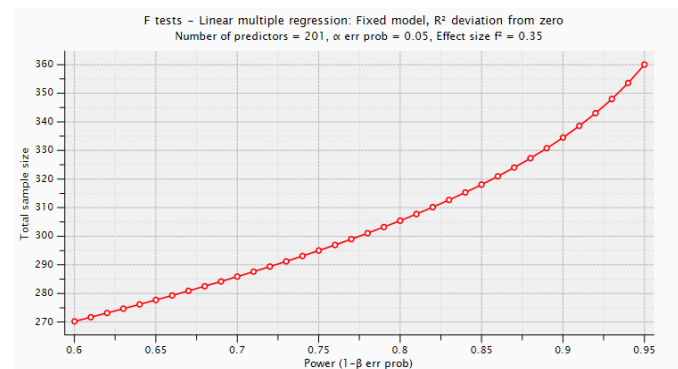


Fig. 1. Results of Power Test to Verify Statistical Significance Level.

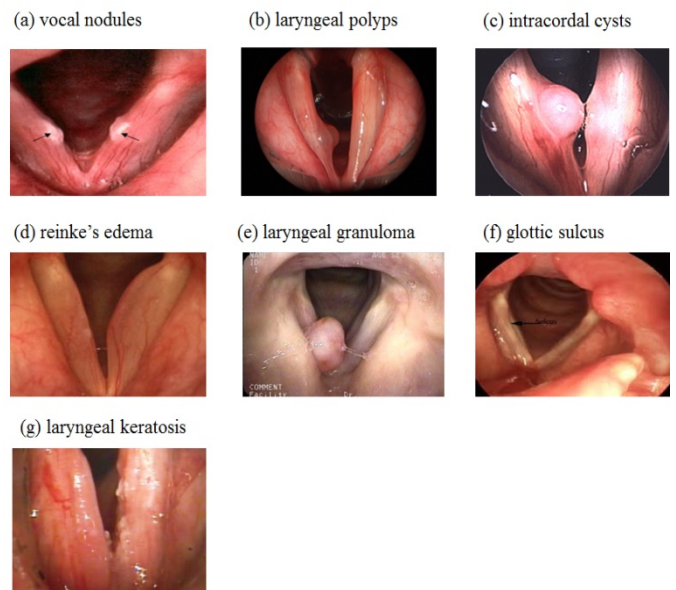


Fig. 2. Type of benign Vocal Fold Mucosal Disorders [38].

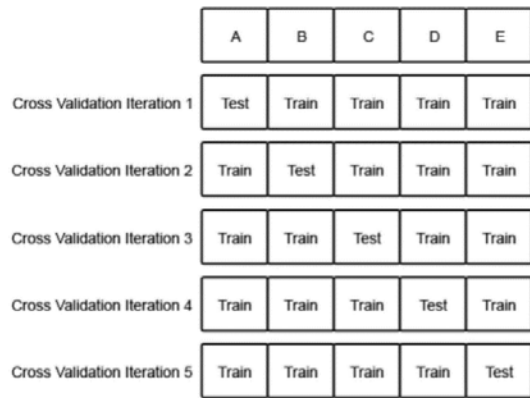


Fig. 3. The Concept of 5-fold Cross-validation.

```

glimpse(data)

#cross validation, using rf to predict sepal.length
k = 5

data$id <- sample(1:k, nrow(data), replace = TRUE)
list <- 1:k

# prediction and test set data frames that we add to with each iteration over
# the folds
prediction <- data.frame()
testsetCopy <- data.frame()

#Creating a progress bar to know the status of CV
progress.bar <- create_progress_bar("text")
progress.bar$init(k)
}
    
```

Fig. 4. R Code for 5-fold Cross-validation.

When developing a model using a method with a random characteristic (e.g., random forest), the seed was fixed to #0123456. This study defined the model with the highest accuracy as the model with the best prediction performance. When the accuracy was identical, a model with a shorter runtime was selected as the model with the best prediction performance. All analyzes were performed using R version 3.6.3 (Foundation for Statistical Computing, Vienna, Austria).

D. Machine Learning Models

The decision tree is an algorithm that creates a learning model in the tree shape according to the features of the data and derives a final decision through repetition. Since the decision tree expresses the analysis process in a tree-shaped graph, the decision tree has the advantage of helping a researcher understand and explain the analysis process easily (Fig. 5). In this study, CART was used as a decision tree algorithm. In this study, the maximum tree depth was set to 10, the parent node was set to 50, and the child node was set to 30.

The naive Bayes model is a method of classifying observations by using Bayes theory (Fig. 6). Bayes theory refers to a way of deriving a posteriori probability for a certain observation by using a secured prior probability.

Random forest is a decision tree-based ensemble method that generates many random samples using a bootstrap (randomly extracting samples of the same size from a given

data with replacement) from a learning data, trains independent decision trees for each sample group, and synthesizes the results to create a final model (Fig. 7).

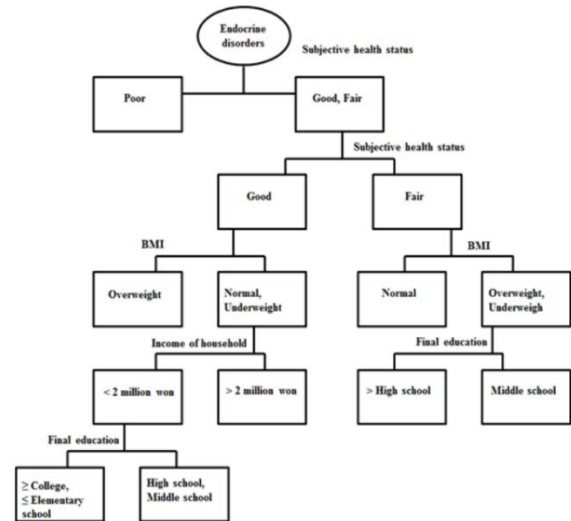


Fig. 5. Example of CART [39].

Input:

Training dataset T,
 $F = (f_1, f_2, f_3, \dots, f_n)$ // value of the predictor variable in testing dataset.

Output:

A class of testing dataset.

Step:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat

Calculate the probability of f_i using the gauss density equation in each class;

Until the probability of all predictor variables ($f_1, f_2, f_3, \dots, f_n$) has been calculated.

4. Calculate the likelihood for each class;
5. Get the greatest likelihood;

Fig. 6. Naive Bayes' Algorithm [40]

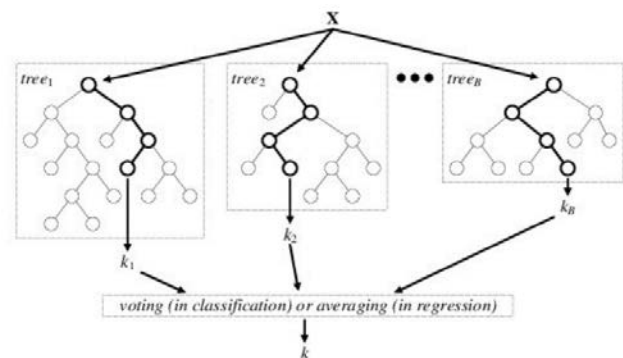


Fig. 7. Concept of Random Forest Model [41].

The generalized linear model is an extension of the linear model that can handle cases where a dependent variable of the dataset does not satisfy the normal distribution assumptions. It is a regression analysis using the glm() function. In other words, the generalized linear model models $f(x)$, which is formed by converting a dependent variable, using a linear combination of the independent variable and the regression coefficient.

Deep learning is an algorithm composed of an input layer, composed of independent variables, an output layer, composed of dependent variables, and two or more hidden layers between the input and output layers. Independent nodes are arranged in each layer, and the nodes between the two layers are connected by weighted neurons (connecting lines) (Fig. 8).

This study used H2O Deep Learning among various deep learning types. H2O's Deep Learning is a type of the multi-layer feedforward artificial neural networks, and it is trained with gradient descent optimization and back-propagation.

In this study, the number of hidden layers was set to 2 (200 hidden node), a default value, and epoch (the number of passes of the entire training dataset) was set to 10. H2O Deep Learning provides Tanh, Tanh with dropout, Rectifier, Rectifier with dropout, Maxout, and Maxout with dropout as activation functions. This study used rectifier, the default function, as an activation function to develop models. The code of H2O Deep Learning is presented in Fig. 9.

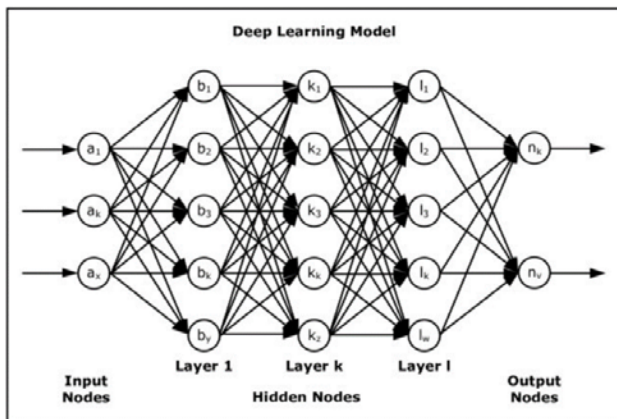


Fig. 8. The Concept of Deep Learning [42].

```
library(h2o)
h2o.init()

y = 'y_2002'
x = set_diff(names(train_data), y)

train.hex = as.h2o(train_data)
new_data.hex = as.h2o(test_data)

dl = h2o.deeplearning(x = x, y = y, training_frame = train.hex, nfolds = 5)

drf = h2o.randomForest(x = x, y = y, training_frame = train.hex, nfolds = 5)

glm = h2o.glm(x = x, y = y, training_frame = train.hex, nfolds = 5)
```

Fig. 9. Code of H2o Deep Learning.

III. RESULTS

A. Comparing the Accuracy and Runtime of benign Laryngeal Mucosal Disorder Prediction Models

The accuracies of five models (deep learning, naive Bayes model, generalized linear model, CART, and random forest) for predicting benign laryngeal mucosal disorders are presented in Fig. 10. In this study, deep learning was the best model with the highest accuracy (0.84). The runtimes of the five models are presented in Fig. 11. In this study, CART showed the shortest runtime (3min 7sec).

B. Predictors of benign Laryngeal Mucosal Disorders in Korean Adults

The normalized importance of the deep learning's variables, the final model, is presented in Fig. 12. This model confirmed that subjective voice problem recognition, pain and discomfort in the last two weeks, education level, occupation, mean monthly household income, high-risk drinker, and current smoker were major variables with high weight for the benign laryngeal mucosal disorders of Korean adults. Among them, subjective voice problem recognition was the most important factor with the highest weight.

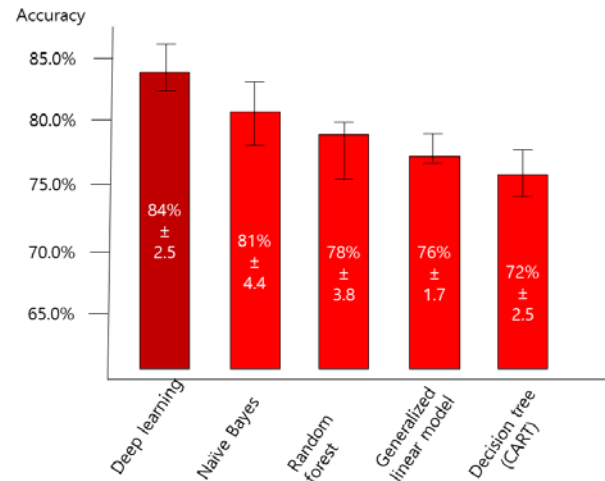


Fig. 10. Accuracy Comparison of Machine Learning and Deep Learning Models for Predicting benign Laryngeal Mucosal Disorders, (%).

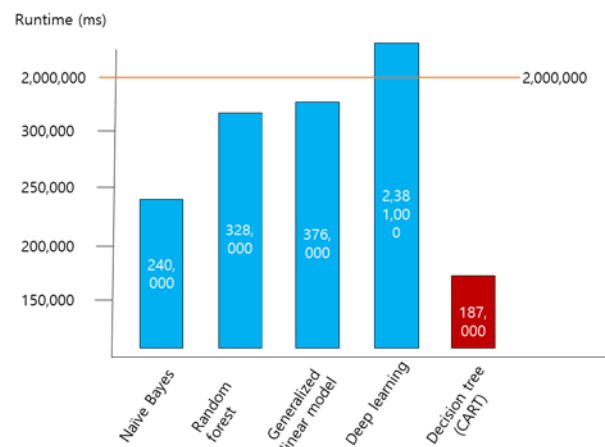


Fig. 11. Runtime Comparison of Machine Learning and Deep Learning Models for Predicting benign Laryngeal Mucosal Disorders, (ms).

Attribute	weight
Self-reported voice problem	0.094
Pain & discomfort over the last two weeks	0.075
Education level	0.066
Occupation	0.043
Average monthly household income	0.038

Fig. 12. Variable's Importance for benign Laryngeal Mucosal Disorders (Only Top 5).

IV. DISCUSSION

This study compared models for predicting the benign laryngeal mucosal disorders of Korean adults. The results of this study showed that deep learning had the best prediction performance among deep learning, naive Bayes model, generalized linear model, CART, and random forest. The runtime of deep learning was 39min 41sec, which was a 10 times longer development time than CART (3min 7sec). However, deep learning showed better ($\geq 6\%$) accuracy than machine learning models. The results of this study agreed with the results of previous studies [36, 43] that reported that the performance of deep learning was better than ensemble-based machine learning methods (e.g., light gradient boosted machine, and extreme gradient boosting) for predicting laryngeal disorders by using video, image, and speech analysis. The results of this study implied that the prediction performance of deep learning could be better than that of machine learning for structured data such as health behavior and demographic factors as well as video and image data. However, since machine learning studies using epidemiological data are much less than machine learning studies using video, image, and speech data, additional studies are needed to prove the superiority of prediction performance of deep learning in epidemiologic data such as health surveys.

V. CONCLUSION

The results of this study suggested that the prediction performance of deep learning could be better than other machine learning methods when developing a multi-modal model for predicting benign laryngeal mucosal disorders by using various data such as image data, demographic factors, and health behavior in the future. It will be necessary to compare the accuracy and runtime of models using the data of various diseases in order to prove the prediction performance of deep learning models, built by using epidemiological data.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07041091, NRF-2019S1A5A8034211).

REFERENCES

- [1] T. M. McCarrel, and J. B. Woodie, Update on laryngeal disorders and treatment. *Veterinary Clinics: Equine Practice*, vol. 31, no. 1, pp. 13-26, 2015.
- [2] H. S. Kim, Benign laryngeal disorders. *Korean Journal of Otorhinolaryngology-Head and Neck Surgery*, vol. 56, no. 6, pp. 332-338, 2013.
- [3] H. D. Soni, S. Gandhi, M. Goyal, and U. Shah, Study of clinical profile of benign laryngeal lesions. *International Journal of Medical Science and Public Health*, vol. 5, no. 4, pp. 656-660, 2016.

- [4] H. Byeon, and Y. Lee, Prevalence and risk factors of benign laryngeal lesions in the adult population. *Communication Sciences & Disorders*, vol. 15, no. 4, pp. 648-656, 2010.
- [5] M. M. Johns, Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts. *Current Opinion in Otolaryngology & Head and Neck Surgery* vol. 11, no. 6, pp. 456-461, 2003.
- [6] H. Byeon, Model development for predicting the occurrence of benign laryngeal lesions using support vector machine: focusing on South Korean adults living in local communities. *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 222-227, 2018.
- [7] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, vol. 115, no. 11, pp. 1988-1995, 2005.
- [8] H. Byeon, and Y. Lee, Laryngeal pathologies in older Korean adults and their association with smoking and alcohol consumption. *The Laryngoscope*, vol. 123, no. 2, pp. 429-433, 2013.
- [9] H. Byeon, Gender differences in risk factors of benign vocal fold disease in Korea: the fifth Korea National Health and Nutritional Examination Survey. *Logopedics Phoniatrics Vocology*, vol. 41, no. 2, pp. 85-91, 2016.
- [10] H. Byeon, A population-based cross-sectional study of alcohol consumption and risk of benign laryngeal disease in Korean adults. *Journal of Voice*, vol. 30, no. 4, pp. 443-447, 2016.
- [11] S. H. Woo, R. B. Kim, S. H. Choi, S. W. Lee, and S. J. Won, Prevalence of laryngeal disease in South Korea: data from the Korea National Health and Nutrition Examination Survey from 2008 to 2011. *Yonsei Medical Journal*, vol. 55, no 2, p. 499-507, 2014.
- [12] H. Byeon, Occupational risks for voice disorders: evidence from a Korea national cross-sectional survey. *Logopedics Phoniatrics Vocology*, vol. 42, no. 1, pp. 39-43, 2017.
- [13] S. Simberg, A. Laine, E. Sala, and A. M. Rönnekaa, Prevalence of voice disorders among future teachers. *Journal of Voice*, vol. 14, no. 2, pp. 231-235, 2000.
- [14] M. Behlau, F. Zambon, A. C. Guerrieri, and N. Roy, Epidemiology of voice disorders in teachers and nonteachers in Brazil: prevalence and adverse effects. *Journal of Voice*, vol. 26, no. 5, pp. 665.e9-665.e18, 2012.
- [15] L. C. C. Cutiva, I. Vogel, and A. Burdorf, Voice disorders in teachers and their associations with work-related factors: a systematic review. *Journal of Communication Disorders*, vol. 46, no. 2, pp. 143-155, 2013.
- [16] H. Byeon, The prediction model for self-reported voice problem using a decision tree model. *Journal of the Korea Academia-Industrial cooperation Society*, vol. 14, no. 7, pp. 3368-3373, 2013.
- [17] P. N. Carding, S. Roulstone, K. Northstone, and ALSPAC Study Team. The prevalence of childhood dysphonia: a cross-sectional study. *Journal of Voice*, vol. 20, no. 4, pp. 623-630, 2006.
- [18] N. R. Williams, Occupational groups at risk of voice disorders: a review of the literature. *Occupational Medicine*, vol. 53, no. 7, pp. 456-460, 2003.
- [19] H. Byeon, Prevalence of perceived dysphonia and its correlation with the prevalence of clinically diagnosed laryngeal disorders: the Korea National Health and Nutrition Examination Surveys 2010-2012. *Annals of Otolaryngology, Rhinology & Laryngology*, vol. 124, no. 10, pp. 770-776, 2015.
- [20] R. M. B. De Alvear, F. J. Barón, and A. G. Martínez-Arquero, School teachers' vocal use, risk factors, and voice disorder prevalence: guidelines to detect teachers with current voice problems. *Folia Phoniatrica et Logopaedica*, vol. 63, no. 4, pp. 209-215, 2011.
- [21] H. Byeon, Relationships among smoking, organic, and functional voice disorders in Korean general population. *Journal of Voice*, vol. 29, No. 3, pp. 312-316, 2015.
- [22] H. Byeon, Comparative analysis of unweighted sample design and complex sample design related to the exploration of potential risk factors of dysphonia. *Journal of the Korea Academia-Industrial cooperation Society*, vol. 13, no. 5, pp. 2251-2258, 2012.
- [23] H. Byeon, Comparing the accuracy and developed models for predicting the confrontation naming of the elderly in South Korea using weighted

- random forest, random forest, and support vector regression. *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 326-331, 2021.
- [24] K. E. Bainbridge, N. Roy, K. G. Losonczy, H. J. Hoffman, and S. M. Cohen, Voice disorders and associated risk markers among young adults in the United States. *The Laryngoscope*, vol. 127, no.9, pp. 2093-2099, 2017.
- [25] L. M. da Rocha, S. de Lima Bach, P. L. do Amaral, M. Behlau, and L. D de Mattos Souza, Risk factors for the incidence of perceived voice disorders in elementary and middle school teachers. *Journal of Voice*, vol. 31, no. 2, pp. 258.e7-258.e12, 2017.
- [26] K. Kyriakou, K. Petinou, and I. Phiniketos, Risk factors for voice disorders in university professors in Cyprus. *Journal of Voice*, vol. 32, no. 5, pp. 643.e1-643.e9, 2018.
- [27] Y. R. Lee, H. R. Kim, and S. Lee, Effect of teacher's working conditions on voice disorder in Korea: a nationwide survey. *Annals of Occupational and Environmental Medicine*, vol. 30, no. 1, pp. 1-10, 2018.
- [28] H. Byeon, D. Lee, and S. Cho, Association between second-hand smoking and laryngopathy in the general population of South Korea. *PLOS ONE*, vol. 11, no. 11, pp. e0165337, 2016.
- [29] H. Byeon, D. Lee, and S. Cho, Relationship between women's smoking and laryngeal disorders based on the urine cotinine test: results of a national population-based survey. *BMJ Open*, vol. 6, no. 11, pp. e012169, 2016.
- [30] H. Byeon, A laryngeal disorders prediction model based on cluster analysis and regression analysis. *Medicine*, vol. 98, no. 31. pp. e16686, 2019.
- [31] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. *Information Sciences*, vol. 474, pp. 106-124, 2019.
- [32] H. Byeon, Evaluating the accuracy of models for predicting the speech acceptability for children with cochlear implantations. *International Journal of Advanced Computer Science and Applications*, vol. 12, pp. 25-29, 2021.
- [33] H. Byeon, Comparing ensemble-based machine learning classifiers developed for distinguishing hypokinetic dysarthria from presbyphonia. *Applied Sciences*, vol. 11, no. 5, pp. 2235, 2021.
- [34] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular Pharmaceutics*, vol. 14, no. 12, pp. 4462-4475, 2017.
- [35] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: comparison with machine learning approaches. *Energies*, vol. 11, no. 7, 1636, 2018.
- [36] H. Xiong, P. Lin, J. G. Yu, J. Ye, L. Xiao, Y. Tao, Z. Jing, W. Lin, M. Liu, J. Xu, W. Hu, Y. Lu, H. Liu, Y. Li, Y. Zheng, and H. Yang, Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine*, vol. 48, pp. 92-99.
- [37] J. Ren, X. Jing, J. Wang, X. Ren, Y. Xu, Q. Yang, L. Ma, Y. Sun, W. Xu, N. Yang, J. Zou, Y. Zheng, M. Chen, W. Gan, T. Xiang, J. An, R. Liu, C. Lv, K. Lin, X. Zheng, F. Lu, Y. Rao, H. Yang, K. Liu, G. Liu, T. Lu, X. Zheng, and Y. Zhao, Automatic recognition of laryngoscopic images using a deep - learning technique. *The Laryngoscope*, vol. 130, no. 11, pp. E686-E693, 2020.
- [38] H. Byeon, Exploring potential risk factors for benign vocal fold mucosal disorders using weighted logistic regression. *International Journal of Bio-Science and Bio-Technology*, vol .6, no. 4, pp. 77-86, 2014.
- [39] H. Byeon, Development of prediction model for endocrine disorders in the Korean elderly using CART algorithm. *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 125-129, 2015.
- [40] M. F. A. Saputra, T. Widiyaningtyas, and A. P. Wibawa, Illiteracy classification using K means-Naïve Bayes algorithm. *JOIV: International Journal on Informatics Visualization*, vol. 2, no. 3, 153-158, 2018.
- [41] A. Verikas, E. Vaiciukynas, A. Gelzinis, J. Parker, and M. C. Olsson, Electromyographic patterns during golf swing: activation sequence profiling and prediction of shot effectiveness. *Sensors*, vol. 16, no. 4, pp. 592, 2016.
- [42] W. Serrano, Smart internet search with random neural networks. *European Review*, vol. 25, no. 2, pp. 260-272, 2017.
- [43] H. Kim, J. Jeon, Y. J. Han, Y. Joo, J. Lee, S. Lee, and S. Im, Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. *Journal of Clinical Medicine*, vol. 9, no. 11, pp. 3415, 2020.

Distance Education during COVID-19 Pandemic: The Perceptions and Preference of University Students in Malaysia Towards Online Learning

Husna Hafiza Razami¹, Roslina Ibrahim²

UTMSPACE, Kuala Lumpur, Malaysia¹

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia²

Abstract—The sudden shift from the brick-and-mortar approach to online distance learning due to the coronavirus pandemic greatly impacted everyone involved, particularly students. Hence, it is critical to identify the perception of students regarding the challenges they faced, their satisfaction with remote learning, as well as their preferences and recommendations for improvement which are the objectives of this research. A survey taken by 408 diploma students with 377 valid answers for the quantitative study showed that the most common difficulties they encountered were in terms of interaction, concentration and motivation. The mean of the perceived challenges was found to be significantly different depending on the respondents' prior e-learning experience and area of residence. With regard to the relevant activities to be conducted virtually, most participants approved of assessments such as quiz, assignments and tests. Animation and gamification received the highest votes as the elements that students wished were incorporated to boost their online learning engagement. The findings from this research contribute to existing studies on the perceptions and preference of students towards distance education by shedding light on the perspective of diploma students.

Keywords—COVID; distance learning; education; online learning; student; perception; preference

I. INTRODUCTION

Distance education is a provision of flexible opportunities due to which students can access formal learning from wherever they are. The three characteristics that define distance education are as follows: 1) the learning community is mostly in a separate space; 2) the process of instruction is assisted by two-way interaction between the instructors and students; 3) technology is used to facilitate learning and communication [1]. In the past decade, many universities worldwide started to offer distance learning programs and courses to reach students who find it difficult to pursue their studies due to constraints of time, geographical distance and commitments.

Various tools have been developed to digitalize educational resources as well as to enable instructors to meet and engage the learners virtually, such as the learning management system, digital books and so on [2]. Despite the increasing adoption of technology, some institutions have not yet implemented online distance learning as they are not prepared for an entirely web-based experience [3], [4]. Online learning is normally considered an alternative or complementary to physical

classroom instruction for gaining knowledge and integrating technology [5], [3].

However, the emergence of the deadly coronavirus disease at the end of 2019, which gradually swept all over the globe, brought instant and dramatic changes in education. As a measure to control and minimize virus transmission, the governments in many countries around the world, including Malaysia, implemented mandatory closure of educational institutions and all their physical teaching and learning activities. In Malaysia, the disease started to emerge at the end of February 2020, and the lockdown was then imposed on the 18th of March, 2020, following the Movement Control Order (MCO) issued by the government. To ensure continuation of learning, academic institutions had no choice but to implement emergency remote education [6].

This abrupt paradigm shift from traditional brick-and-mortar learning to completely internet-based instruction greatly impacted everyone involved, particularly the students who registered for face-to-face education in campus, but, all of a sudden, had to switch from the conventional style of learning that they desired and had been accustomed to. Further, those residing in rural areas or in a non-conducive environment face greater challenges to learning efficiently online [7]. Educators and learning institutions also encountered difficulties in adapting to the sudden change [6], [8]. For instance, lecturers have to tailor activities and assessments to the digital environment and also had to learn to use various technology-enhanced educational tools in a short span of time. Moreover, unstable or insufficient network also poses unprecedented challenges for both lecturers and students when conducting live video conferencing sessions [9].

School of Professional and Continuing Education (SPACE), a faculty of Universiti Teknologi Malaysia (UTM), had adopted Blackboard as their new learning management system before the COVID-19 pandemic. The adoption was still at the infant stage: Not everyone in the community utilized the system, and the basic training, which was also ongoing, primarily focused on the features of the technology that complement classroom education. When campus closure was enforced by the government, training for online learning using Blackboard was given to all lecturers and students, but lecturers were allowed to use any platform or tool they preferred or considered as a better fit for their course and their students' preference as long as the educational activities were

conducted as per schedule and the learning outcomes were achieved. This flexibility was allowed to minimize the stress and disruption caused by the coronavirus outbreak.

Hence, during remote teaching, Blackboard was the official platform used by the majority of the lecturers in SPACE, UTM, followed by Google Classroom and other mediums for asynchronous activities that involved no real-time interaction. In terms of synchronous activities that required concurrent two-way communication between lecturers and students, the commonly used video conferencing platforms were Blackboard Collaborate, Google Meet, Cisco Webex and Zoom.

More than a year has passed, but the situation is still unstable, with the COVID-19 case numbers fluctuating and sometimes substantially rising. This has forced universities and faculties in Malaysia, including SPACE, to remain primarily closed. The ambiguous future demands that governments and learning institutes plan and implement a flexible and robust education system. Considering this and the momentous acceleration of digital learning due to the pandemic, it can be expected that the Internet will continue playing a significant role in human life. SPACE, for example, is currently preparing to develop Open and Distance Learning (ODL) programs and micro-credentials. Thus, evaluating the current adoption of remote learning as well as the students' experiences of and preference towards online education is critical for providing guidance to institutions and assisting them in deciding on, designing and implementing a relevant and fulfilling approach to education [9].

Therefore, the objectives of this research are to 1) determine students' perceptions of the online distance education conducted during the COVID-19 pandemic; 2) identify students' preference for online learning. Even though various aspects, such as challenges, implications and strategies, have been discussed and studied by many researchers [5], [10], [13], [25], most of them focused on a certain subject or presented a general view of the students' perception despite their diverse backgrounds. Hence, this research intends to enrich the current literature by providing the perspective of diploma students, which, to the best of the author's knowledge, has not yet been discussed. Most of the prior studies collected data mainly from postgraduate, degree and foundation students [10]. Furthermore, the findings on the difference of students' perception among gender, household income, prior e-learning experience and area of residence can be useful for institutions when planning the online teaching and learning activities. The following sections of this paper comprise of the literature review, methodology, results and discussion, followed by conclusion and future works.

II. LITERATURE REVIEW

A. Students' Perceptions of Online Learning

Investigations had been carried out in the form of numerous studies to determine students' perceptions and attitudes towards the completely digital education during the coronavirus outbreak, and one of the popular topics discussed was the challenges faced. For example, [11] found that the biggest concern among students in Jordanian medical schools was poor Internet coverage, followed by Internet data

limitation, learning platform variation and insufficient devices. Similarly, the findings of another study showed that 69% of 762 students from two universities in Romania faced technical problems with the platform used by the university, whereas 14.8% had inadequate devices and mobile connection. Other challenges that the learners had to cope with were mainly connected to the lecturers, such as limited employment of tools, lack of necessary skills and motivation to improve and adapt their teaching styles to the digital environment, provision of insufficient support for easing students' learning process as well as the giving of an excessive number of tasks and poor adherence to the learning schedule [12].

Another research summarized and divided the learning barriers encountered by 670 medical students into five categories: technological issues, individual issues (such as health and challenges to adjust to the learning style), domestic issues (such as a non-conducive environment and family and financial problems), institutional barriers and community barriers [13]. In the context of Malaysia, 147 students from 16 universities expressed issues related to administration, social interactions, technical problems, Internet connection, learning time, learning support and motivation [7]. Lack of interaction was also pointed out in another research [5], [10]. Generally, the results from these studies indicated that students encountered similar difficulties regardless of their nationality and background, but differences also existed.

On the positive side, students do acknowledge the advantages of e-learning, such as being able to continue their education remotely from the comfort of their homes, the usefulness of having recorded videos to facilitate their studying and revision process and the development of their self-directed learning skills [14]. Results from other literature showed comfortable learning environment as one of the most prevalent benefits of online education in addition to the flexibility of learning anytime and anywhere and the reduction of transportation time and expenses [5], [11]. For enhancing online learning, students suggested that the interaction be improved and the workload be reduced [14]. Nevertheless, some studies discovered the students' satisfaction with the e-learning classroom interactions as well as their academic achievement during the pandemic [5].

B. Students' Preference of Online Learning

One of the frequently asked questions in past researches on learners' preference with regard to digital instruction was related to the mode of learning: Do learners favor a completely online, a completely face-to-face or a blended, hybrid approach? Blended was the most desired method according to multiple studies [15], [16], [17]. Although students approved of the benefit of learning flexibility that comes with online education, they were not keen on the idea of shifting entirely to the digital mode as they did not wish to lose the human aspect and the collaborative opportunities that come from physical interaction, which they believed were crucial for adapting to the workplace environment in the future [18].

The author in [24] showed that agriculture students were more inclined towards learning via smartphone followed by laptop compared to other devices and through recorded videos rather than live sessions or reading materials. In terms of class

schedule and class duration, the findings revealed that learners wished to have 45-minute classes twice per week, with 15-minute breaks between classes. The majority of the survey respondents indicated that they wanted the course material to be complemented with videos that have instructors explaining using whiteboard or PowerPoint. In terms of assessment, 76% of them wished to have quizzes and assignments at the end of each class for effective learning and academic success. They also preferred to be given one week to submit assignments or assigned a due date that was before the next scheduled class session. The students stated that they liked attending online examinations, particularly if they had multiple-choice-question format [18].

III. METHODOLOGY

A. Participants and the Scope of the Study

This research was carried out in SPACE, a faculty of UTM, a higher education institution in Malaysia. From a population of 1347 full-time diploma students, those undergoing industrial internship were excluded, and 408 (30%) students from all 18 diploma programs offered by the university participated in the survey. The students belonged to the engineering, management, computer science and services or geomatics and built environment departments. None of them had experienced online distance learning programs or courses before the pandemic. They are local students who had enrolled at the university expecting to receive formal education on campus, but since March 2020, they had to attend classes online from their homes due to the movement control order implemented by the Malaysian government. This study was conducted from the 27th of December, 2020, to the 10th of January, 2021; by this time, the students had experienced eight months of fully online learning. Blackboard was the main learning management system used by the students, and other applications such as WhatsApp, Google Classroom and many more were also used.

During data cleaning, 31 responses had to be removed for the Likert scale questions' quantitative analysis, leaving 377 (28%) responses. The reason behind this elimination is explained in the next section, under data collection and analysis. Table I shows the demographic characteristics of the 377 respondents, which comprised of students from semesters one through six.

Overall, majority of the students live in urban areas (76.9%), and used their home Wi-Fi to access the Internet for their e-learning activities during the pandemic (78.2%). Almost 90% respondents had moderate-level information technology (IT) skills. As for the time spent every day for online learning, 237 (62.9%) of them chose five to eight hours, which is the expected average daily learning time based on the students' learning schedules. Students' household income was classified into three categories: B40 for the low-income category (less than RM 4,849), M40 for the middle-income category (between RM 4,850 and RM 10,959) and T40 for the high-income category (RM 10,960 or more).

B. Data Collection and Analysis

This research employed a cross-sectional survey in which the data was collected at the same time [19]. Purposeful

sampling strategy was opted for because the scope of this research was diploma students from the same university who had experienced online distance learning for the first time. The online questionnaire used for this study began with an introduction to brief the respondents about the purpose of the survey and assure them about their anonymity and the confidentiality of their responses. The questionnaire consisted of three main segments: demographic information, students' perception and students' preference, using multiple choice questions, Likert scale and open-ended questions developed based on relevant literature and modified to suit the context of this research.

The items used to measure satisfaction were adapted from [20], and the items for measuring online learning advantages and challenges were adapted from [18] and [21]. The survey questions were written in English as well as Bahasa Melayu to ensure that no language barrier problem occurred if any of the respondents did not understand the former language well. Since the respondents came from various programs of study in which their respective lecturers employed different tools, general questions were used for the survey; no questions were asked about specific subjects or learning platforms.

Once the questionnaire was developed in Google Forms, it was reviewed by an educational technology expert and assessed through a pilot test. The findings from the pilot test indicated good reliability of the Likert scale questions based on Cronbach's alpha values (0.854 for satisfaction, 0.881 for usefulness, 0.760 for challenges). Nonetheless, some amendments were done based on students' comments on questions that they found unclear and difficult to understand. The improved questionnaire was then sent to all respective heads of programs with the request to share it with their students via WhatsApp. This approach was selected to ensure that feedback was obtained for each program offered by the university. An e-gift voucher was emailed to them as a token of appreciation for supporting this research.

The collected data was then examined, first based on the five-point Likert scale questions to detect straight-lining and contradictory answers as they represent responses that were not carefully and thoughtfully given. Such responses need to be removed as they can reduce the validity of the quantitative study [22]. However, the feedbacks obtained for open-ended questions were all considered for analysis.

Once the data was screened, the quantitative results were analyzed using the SPSS software for descriptive and inferential statistics, including independent samples t-test, and Spearman's correlation analysis. On the other hand, the analysis of the open-ended feedbacks was based on the constant comparative method, which involves data coding followed by the categorization and comparison of data to identify the themes for similar responses and, finally, the calculation of the frequency for each theme [23].

IV. RESULTS AND DISCUSSION

A. Students' Perceptions of the Online Learning Conducted During the Coronavirus Pandemic

The way students viewed the remote education conducted during the pandemic was studied under three categories. The

first one was regarding their satisfaction, and the results from six Likert scale questions, as displayed in Table II, revealed not very impressive contentment among the respondents.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF 377 STUDENTS

Characteristic	Category	N	Percentage
Gender	Male	170	45.1%
	Female	207	54.9%
Internet access used for learning	Home Wi-Fi	295	78.2%
	Mobile hotspot	179	47.5%
Information technology skills level	High	22	5.8%
	Moderate	326	86.5%
	Low	29	7.7%
Have prior online learning experience	Yes	65	17.2%
	No	312	82.8%
Current area of residence	Rural	87	23.1%
	Urban	290	76.9%
Daily hours spent on e-learning	Less than 5 hours	30	7.9%
	5 to 8 hours	237	62.9%
	9 to 12 hours	71	18.9%
	More than 12 hours	39	10.3%
Household income category	B40 (Low income)	157	41.6%
	M40 (Middle income)	161	42.7%
	T40 (High income)	59	15.6%

TABLE II. DESCRIPTIVE STATISTICS FOR ONLINE LEARNING SATISFACTION

Item	SD (%)	D (%)	N (%)	A (%)	SA (%)	Mean
I am satisfied with the online classes conducted during this pandemic	8.5	16.4	36.3	30.0	8.8	3.14
I am satisfied with the online learning materials provided by lecturer	3.2	9.3	28.4	43.5	15.6	3.59
I am satisfied with the online assessments such as assignment, quiz, test and exam	4.2	7.7	29.7	38.5	19.9	3.62
I am satisfied with lecturer's teaching delivery through online platform	5.0	12.7	31.6	34.2	16.4	3.44
I am satisfied with the gained knowledge and skills	5.8	12.2	34.0	35.8	12.2	3.36
I am interested to continue learning through online medium	34.5	15.6	25.2	15.4	9.3	2.49

SD: strongly disagree, D: disagree, N: not sure, A: agree, SA: strongly agree

The most negative outlook was observed in students' intention to continue learning through the online medium, with not more than a quarter of them showing interest, and the mean score for it was also the lowest compared to that for the other items. Previous researches received similar results: They showed that more than 60% of the respondents did not plan to continue using e-learning in the future [21], [24]. Another study revealed that students preferred the semester to be postponed and the digital education to be halted during the pandemic [25].

One of the factors that might have caused this discouraging emotion among students was probably the emergency implementation of distance learning, which went against their will, habit and expectation to learn on campus. This aspect is further explored in the next section, which is on the preference of students towards digital education.

For the first five items in the satisfaction category, the total percentage of students who agreed or strongly agreed was more than the percentage of those who showed disapproval, particularly for items two, three and four, with about half of the participants indicating their satisfaction with the learning materials, assessments and online platform. In order to understand the reasons behind the displeasure of the participants, the challenges encountered by the students were discussed and their e-learning experience was further explored in the next category of the questionnaire. It is interesting to note that another research that found positive satisfaction among respondents also obtained contradicting results for the intention to continue e-learning [23].

Given the sudden change from learning on campus to learning remotely, which may have caused distress among students, the participants were asked to rate the perceived usefulness of online education through 11 questions, as showcased in Table III, before being surveyed on the challenges and preferences. This was done to make them realize that this learning method also has its advantages. Overall, mixed results were observed: Based on the mean score, average feedbacks from students ranged from 2.77 to 4.11. The highest ones were not a surprise as the provision to access learning materials anytime from anywhere and re-watch recordings were widely acknowledged benefits of virtual instruction.

Apart from that, more participants agreed that digital learning is useful because they found it flexible, it did not require them to travel to attend lectures, the money spent was less than that spent when learning on campus, they felt comfortable attending online classes and learning online, they could improve their independent and technical skills and they were able to get help from lecturers, who also responded quickly. However, the opposite result was obtained for interaction: A slightly higher percentage of respondents found interacting virtually with lecturers and students uncomfortable. This finding is in line with the multiple studies that highlighted social interaction as one of the barriers to remote education; thus, this element needs to be looked into when designing online courses to improve the impression of students with regard to distance education and enhance their learning experience [5], [7], [10].

TABLE III. DESCRIPTIVE STATISTICS FOR STUDENTS' PERCEPTION OF THE BENEFITS OF ONLINE LEARNING

Item	SD (%)	D (%)	N (%)	A (%)	SA (%)	Mean
I do not need to travel to campus to attend online class	15.1	9.0	20.4	23.3	32.1	3.48
The expenses of online learning are less than those of learning on campus	5.8	7.7	23.9	28.1	34.5	3.78
Online classes are flexible	8.2	9.5	36.1	27.9	18.3	3.38
I can re-watch video recording	2.4	4.5	13.0	30.5	49.6	4.20
I can access learning materials anywhere and anytime	1.1	4.0	18.6	35.8	40.6	4.11
It is comfortable to attend online lectures and learn online	14.9	16.2	31.8	22.0	15.1	3.06
It is comfortable to interact with lecturers through online medium	15.4	20.7	36.9	17.8	9.3	2.85
It is comfortable to interact with classmates through online medium	20.7	20.2	31.3	17.5	10.3	2.77
Online learning allows me to learn to be independent	10.9	8.0	31.6	27.3	22.3	3.42
Online learning allows me to improve my technical skill in using electronic gadgets	3.4	5.8	24.4	38.2	28.1	3.82
I can ask lecturers questions and receive a quick response through online medium	12.7	15.6	31.6	28.4	11.7	3.11

SD: strongly disagree, D: disagree, N: not sure, A: agree, SA: strongly agree

Table IV summarizes the descriptive statistics for students' responses with regard to the challenges they experienced when learning remotely during the COVID-19 pandemic; this section comprised of eight questions. With regard to the first and second items, more than half of the respondents had difficulty interacting with their lecturers and classmates, proving the result obtained earlier. On the other hand, quite a neutral perception was displayed with regard to the problem of Internet connection, perhaps because nearly 80% respondents live in urban areas. A slightly higher percentage of participants struggled with poor learning conditions, lack of self-discipline and adjusting their style of learning compared with those who did not have problems with these.

The most prominent obstacles faced by students were difficulty in staying focused and lack of motivation. Therefore, instructors can try to conduct more engaging learning activities and think of ways to sustain the attention of students during lessons. Lecturers, academic advisors and counselors need to be more attentive and help boost students' motivation more often.

TABLE IV. DESCRIPTIVE STATISTICS FOR STUDENTS' PERCEPTION OF THE CHALLENGES OF ONLINE LEARNING

Item	SD (%)	D (%)	N (%)	A (%)	SA (%)	Mean
It is difficult to interact with lecturer through online medium	6.1	12.5	25.7	28.4	27.3	3.58
It is difficult to interact with classmates through online medium	7.7	13.3	22.0	27.3	29.7	3.58
I have poor, limited and unstable internet connection, which affects my online learning	12.5	21.2	29.4	19.9	17.0	3.08
I have poor learning conditions at home, which affects my online learning	9.5	18.3	28.6	23.3	20.2	3.26
I have lack of self-discipline, which affects my online learning	9.0	19.4	31.8	23.9	15.9	3.18
It is difficult to adjust my learning style	7.4	19.4	32.9	23.1	17.2	3.23
It is difficult for me to stay focused during online learning	3.2	7.2	25.5	27.3	36.9	3.88
I have less motivation when learning online compared to face-to-face learning	3.7	8.0	24.1	29.2	35.0	3.84

SD: strongly disagree, D: disagree, N: not sure, A: agree, SA: strongly agree

B. Differences in Students' Perceptions based on Gender, Household Income, Online Learning Experience and Area of Residence

According to the independent samples t-tests' results, there was no significant difference in the mean of satisfaction, perceived usefulness and challenges based on gender and household income. Nevertheless, students who had prior online learning experience had lower perceived challenges ($M = 3.22$, $SD = 0.98$) than those who did not have experience ($M = 3.50$, $SD = 0.79$), $t(82.542) = -2.161$, $p = 0.034$. The t-test analysis for each item showed that the significant difference was in terms of the difficulty to adapt learning styles $t(375) = -3.157$, $p = 0.002$, as well as remain focused $t(375) = 3.738$, $p = 0.000$, and motivated $t(82.979) = -3.028$, $p = 0.003$.

Students who lived in rural regions demonstrated substantial difference in their perceptions of the challenges during the emergency remote education $t(375) = 2.336$, $p = 0.02$, specifically in terms of having poor learning conditions at home $t(375) = 4.195$, $p = 0.000$, in addition to poor, limited and unstable internet connection, which they felt were affecting their online learning $t(375) = 3.515$, $p = 0.000$. Nevertheless, there was not much difference in the overall satisfaction

between respondents from rural areas and those from urban areas $t(375) = 0.595$, $p = 0.552$, and this gave the impression that the area of residence factor did not play a critical role in satisfying online learners. The finding of this study also reveals that participants' gender, digital learning experience, area of residence and whether their family have low income or not did not influence their perceptions of the advantages of online learning.

C. Relationship of Semester of Study, IT Level and Household Income with Students' Perceptions

This research investigated the association of several factors with the perceptions of students using Spearman's correlation analysis. Respondents' semester of study exhibited statistically significant weak negative correlation with distance learning satisfaction ($\rho = -0.255$, $p = 0.000$) and perceived usefulness ($\rho = -0.212$, $p = 0.000$), and it exhibited a weak positive relation with the challenges encountered ($\rho = 0.207$, $p = 0.000$). This might be because students in the higher semesters need to learn more difficult subjects that require them to take up more hands-on and practical activities to gain a better understanding and develop relevant skills. Therefore, the limitations of remote learning might pose more difficulties and give more disappointment to them compared with students in lower semesters of study.

On the other hand, the associations of the IT level with satisfaction ($\rho = 0.057$, $p = 0.273$) and perceived usefulness ($\rho = 0.071$, $p = 0.172$) were found to be insignificant. In contrast, there was a significant weak negative correlation between the IT level and the challenges faced by participants during the emergency remote education ($\rho = -0.111$, $p = 0.031$), indicating that students with higher-level IT skills encountered less obstacles when learning through the online medium.

Prior independent samples t-tests results showed that students' family status (whether low income or not) produced insignificant difference in students' perceptions towards digital learning. This correlation analysis further proved that there was indeed no significant relation between household income ($\rho = 0.087$, $p = 0.092$) and satisfaction ($\rho = 0.087$, $p = 0.092$), perceived usefulness ($\rho = 0.065$, $p = 0.211$) or challenges experienced by students when learning remotely during the pandemic ($\rho = -0.036$, $p = 0.483$).

D. Students' Preference of Online Learning

Students' preference towards technology-enhanced education was analyzed based on all 408 original responses. When the respondents were asked about their preferred future learning method, blended format, which is a combination of both online instruction and physical instruction, received the highest votes from the participants; 167 (40.9%) respondents selected this approach, as illustrated in Fig. 1. This outcome, which is consistent with the discoveries in prior studies [15], [16], [26], demonstrated the learners' desire to have technology embedded into their learning experience and also retain face-to-face interaction. In contrast, a study that investigated the perspective of students from education programs found out that the respondents favored face-to-face courses over other types of learning courses [3].

As shown in Fig. 2, the students chose assessments as the most relevant activities to be conducted via the digital platform, particularly quiz, followed by test and assignment. Since their online assessments were basically conducted in an open-book format, students might have felt more relaxed and less stressful in this mode compared with the traditional assessment mode they were used to. A survey in a prior study also received similar results; 60% of the 307 participants revealed that they liked attending online examinations [18]. On the contrary, another study revealed that for medical and nursing students, tests and examinations were the least preferred activities, whereas lectures and discussion were more appropriate [16]. Hence, the preferences of students can be said to differ depending on the type of program or course they take.

For other activities, half of the students felt that discussion could be carried out via the online medium, whereas only 24% of them thought that a group project was viable. Given the advanced technology and the various communication tools present in this digital age, such as social media, discussion forums and live video conference platforms, it was not a surprise that students had no problem in conversing with their peers online. Nonetheless, that a low number of students approved of online group project was understandable, given the limitations to having productive and effective collaboration and team activities virtually.

As shown in Fig. 3, the most popular online educational material was, understandably, digital notes, with 78% (319 out of 408 participants) choosing it. Video tutorial was the next most favored, being approved by almost 73% respondents. This finding was consistent with a previous study in which 84% of 307 students wanted their reading material to be supplemented with video content [18]. Moreover, since these are also the two most prominent features of distance learning courses such as Massive Open Online Courses, this result was probable [27].

In terms of the elements that students thought could make digital education engaging, as demonstrated in Fig. 4, animation and game-based learning were the top two most-favored components, followed by virtual reality activities and simulation. As engagement has been shown to have an impact on making learning easier for students and indirectly influence their attitude towards the acceptance of an online learning tool [27] instructors as well as online course designers and developers can incorporate the aforementioned elements to promote students' participation and enjoyment in learning.

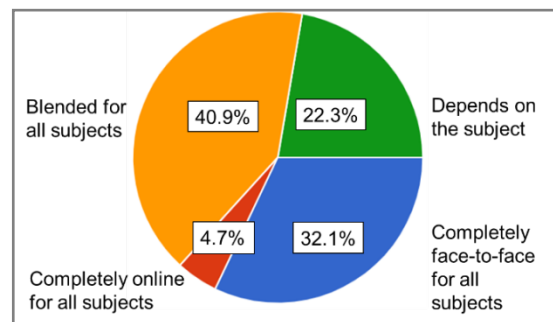


Fig. 1. Percentages for Students' Preferred Future Learning Methods (N = 408).

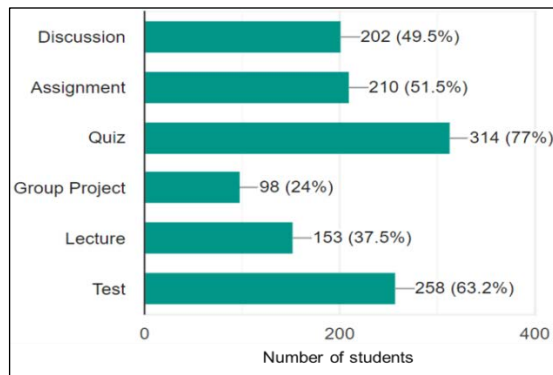


Fig. 2. Activities Recommended by Students as Suitable for Conducting online (N = 408).

Since online education includes both synchronous and asynchronous activities, it is interesting to determine the preference of students regarding the delivery of lecture to support the reading materials. From the responses obtained as displayed in Fig. 5, more than half of the participants chose a combination of live video conferencing and pre-recorded video.

Both methods have their own advantages and limitations, for instance, live session allows real-time interaction but uses more internet data compared to other learning activities. On the other hand, pre-recorded videos provide the benefit for learners to receive and digest content in small chunks. Therefore, taking this feedback from students into consideration, educators can record several short videos for each chapter to let students go through before the live session which may just be spent on discussing certain parts that need further clarification.

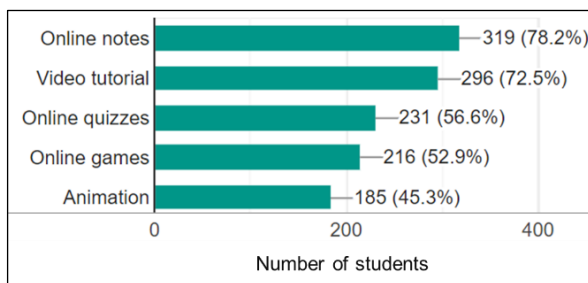


Fig. 3. Learning Materials that Students Preferred (N = 408).

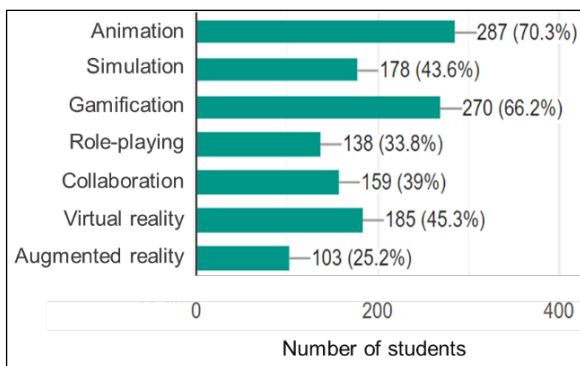


Fig. 4. Number of Students who Approved the Elements can make their e-learning Engaging (N = 408).

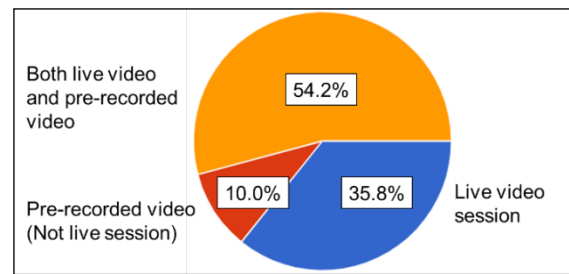


Fig. 5. Percentages of Students According to their Preferred Method (between Live Video and Pre-recorded Video).

The survey questionnaire also included an open-ended question asking for students' suggestions for improving their online learning. The responses were coded and classified into categories, as presented in Table V, which include feedbacks that could be interpreted and found to be relevant to the question. The responses that were not counted mostly said that they (the respondent) had no suggestion and that everything was good.

Recommendations related to teaching style and learning activities were mostly about giving students short breaks between sessions and making classes fun and creative, for instance, through incorporation of online games or more interaction with students, so that they will not feel bored and sleepy. This is consistent with the findings displayed in Fig. 4, which shows that more than 60% students felt that game-based activities can make their learning attractive. Lecturers were suggested to conduct different kinds of activities in class, instead of just lecturing, to boost students' concentration, engagement and motivation.

Some of the respondents also emphasized that lecturers should give clear and thorough explanation using bilingual instruction, provide learning materials before the live session and respond quickly when needed. Apart from that, participants wanted instructors to enhance digital literacy and become more proficient at using the learning platforms as well as reduce group works due to the difficulty in discussing with teammates virtually.

TABLE V. FREQUENCY DISTRIBUTION OF THE SUGGESTIONS FOR IMPROVING ONLINE LEARNING AMONG 408 STUDENTS

Suggestion category	N	Percentage
Improve teaching style and learning activities	71	17.4%
Allow entry to campus	63	15.4%
Others	20	4.9%
Provide free, good and unlimited Internet access	19	4.7%
Improve learning materials or content	18	4.4%
Reduce workload	11	2.7%
Reduce live video session duration	11	2.7%
Implement hybrid or blended learning	6	1.5%
Use platform that consumes less data	6	1.5%
Provide longer assessment duration	5	1.2%
Pre-record short video lesson	5	1.2%

Another very common response from students was regarding their desire to return to the campus. One respondent wrote, "I like to learn through physical class more as it is easier for me to ask questions and I can also understand faster." Another student stated that when they needed to learn at home, challenges such as family problems and family commitments hindered their learning. Other participants who expressed similar issues conveyed that since they needed to do household chores, they hoped that the lecturers could reduce the workload given to them.

Due to these problems, some participants suggested that the university should allow them to enter campus even if learning activities continued to be conducted online, so that they could focus better on their studies and work without disturbance. In contrast, other respondents stressed that remote learning be continued until there are no more COVID-19 cases; such feedbacks were included under the 'Others' category. The differences in students' opinions probably resulted from the dissimilar learning environments they had at their homes. Some participants also recommended hybrid instruction. One respondent said, "If can make it hybrid like physical learning for some days while the rest is online," and another student suggested the same for subjects that were difficult or required practical activities to be carried out on campus.

In terms of live sessions, one participant proposed, "Reduce the duration of learning to 40 minutes as the eyes become painful staring at the screen for too long." Another respondent shared, "It could be hard for us to stay focused and sit in one spot for a long period if the lecture is about two hours." In line with this, some students recommended that instructors pre-record short videos as they thought that these would be more useful and easier to refer to later compared with the recorded long-duration live video sessions. Moreover, respondents also wished that their lecturers prepared easy-to-understand, effective, comprehensive, complete and interesting learning material. Since online notes are the most desired educational material, as shown in Fig. 3, lecturers can enhance the quality of the notes they provide to students by following their suggestions.

V. CONCLUSION

The findings of this study showed that the satisfaction of the respondents was generally neutral to positive. Among the challenges encountered by them, the most common were connected with social interaction, concentration and motivation in learning. There was significant difference in the mean of the participants' perceived challenges based on their previous online learning experience. Blended format was revealed to be the most preferred learning method among respondents with animation and gamification receiving the highest vote for the elements that they believe can engage them.

Even though the research on the distance education during the coronavirus pandemic has rapidly grown, to the best of the authors' knowledge, the perspective of diploma program students has not yet been investigated. This research also evaluated the difference in students' perception based on their household income, area of residence, e-learning experience, semester of study and IT level. Hence, results from this study can enrich the existing literature on the learners' opinions on

remote online learning. Furthermore, educational institutions, instructors as well as online course developers and designers can benefit from the findings of this research; they can understand the current situation, the attitudes of students towards digital instruction and their preferences and suggestions concerning this modern learning method. For instance, the majority of the students preferred a hybrid approach in which the content is delivered online through both live video sessions and pre-recorded videos. The most desired learning materials were online notes followed by video tutorials, quizzes, games and animation. Due to this pandemic, many institutions may have started planning for distance learning programs and courses. Therefore, the results presented in this study can assist them in designing and developing courses that can fulfill the needs and expectations of the students.

Since the findings of this study were obtained from diploma program students in a particular university, the results are relevant for similar cases but cannot be generalized to represent all higher education students. Future research may seek to include students from various institutions and program of study and evaluate the perception among them. The factors that influence learners' satisfaction towards distance education can also be investigated to assist in the continuous quality improvement of this learning method.

ACKNOWLEDGMENT

Authors thank UTMSPACE for the research opportunity under UTMSPACE Research Grant: Potential Development Fund SP-PDF2003 as well as Rozana Ismail, Erni Syuhada Mazwil Ishan, Chin Wei Bing, Nur Anis Syakira Awang and Puteri Sofia Mohd Noor for their assistance during the data collection and analysis.

REFERENCES

- [1] B. D. M. Casey, "The historical development of distance education through technology," *TechTrends: Linking Research and Practice to Improve Learning*, vol. 52, no. 2, pp. 45-51, Apr. 2008. DOI: 10.1007/s11528-008-0135-z, [Online].
- [2] F. Martin, B. Stamper, and C. Flowers, "Examining student perception of readiness for online learning: Importance and confidence," *Online Learn. J.*, vol. 24, no. 2, pp. 38-58, 2020. DOI: 10.24059/olj.v24i2.2053, [Online].
- [3] C. Coman, L.G. Țiru, L. Meseșan-Schmitz, C. Stanciu, and M. C. Bularca, "Online teaching and learning in higher education during the coronavirus pandemic: Students' perspective," *Sustainability (Switzerland)*, vol. 12, no. 24, pp. 1-22, 2020. DOI: 10.3390/su122410367, [Online].
- [4] K. Thompson and J. M. Lodge, "2020 vision: What happens next in education technology research in Australia," *Australas. J. Educ. Technol.*, vol. 36, no. 4, pp. 1-8, 2020. DOI: 10.14742/ajet.6593, [Online].
- [5] N. Arifiati, E. Nurkhayati, E. Nurdiawati, G. Pamungkas, S. Adha, A. Purwanto, O. Julyanto, and E. Azizi, "University students online learning system during Covid-19 pandemic: Advantages, constraints and solutions," *Sys. Rev. Pharm.*, vol. 11, no. 7, pp. 570-576., 2020 [Online]. Available: <http://www.sysrevpharm.org/?mno=9626>.
- [6] L. M. Hasani, H. R. Adnan, D. I. Sensuse, Kautsarina, and R. R. Suryono, "Factors affecting student's perceived readiness on abrupt distance learning adoption: Indonesian higher-education perspectives," 2020, pp. 286-292. DOI: 10.1109/ic2ie50715.2020.9274640.
- [7] A. Ilias, N. Baidi, E. K. Ghani, and F. M. Razali, "Issues on the use of online learning: An exploratory study among university students during

- the COVID-19 pandemic," *Univ. J. Educ. Res.*, vol. 8, no. 11, pp. 5092-5105, 2020. DOI: 10.13189/ujer.2020.081109, [Online].
- [8] F. Alturise, "Difficulties in teaching online with blackboard learn effects of the COVID-19 pandemic in the western branch colleges of Qassim University," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 74–81, 2020, doi: 10.14569/IJACSA.2020.0110512.
- [9] J. K. M. Sia and A. Abbas Adamu, "Facing the unknown: Pandemic and higher education in Malaysia," *Asian Educ. Dev. Stud.*, vol. ahead-of-print, no. ahead-of-print, 2020. DOI: 10.1108/AEDS-05-2020-0114, [Online].
- [10] A. A. Kamal, N. M. Shaipullah, L. Truna, M. Sabri, and S. N. Junaini, "Transitioning to online learning during COVID-19 Pandemic: Case study of a Pre-University Centre in Malaysia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 217–223, 2020, doi: 10.14569/IJACSA.2020.0110628.
- [11] M. Al-Balas, H. I. Al-Balas, H. M. Jaber, K. Obeidat, H. Al-Balas, E. A. Aborajooh, R. Al-TaHER, and B. Al-Balas, "Distance learning in clinical medical education amid COVID-19 pandemic in Jordan: Current situation, challenges, and perspectives," *BMC Med. Educ.*, vol. 20, no. 1, pp. 1-7, 2020. DOI: 10.1186/s12909-020-02428-3, [Online].
- [12] R. E. Baticulon, N. R. I. Alberto, M. B. C. Baron, R. E. C. Mabulay, L. G. T. Rizada, J. J. Sy, C. J. S. Tiu, C. A. Clarion, and J. C. B. Reyes, "Barriers to online learning in the time of COVID-19: A national survey of medical students in the Philippines," *MedRxiv*, pp. 1-19, 2020. DOI: 10.1101/2020.07.16.20155747, preprint.
- [13] K. Mukhtar, K. Javed, M. Arooj, and A. Sethi, "Advantages, limitations and recommendations for online learning during COVID-19 pandemic era," *Pak. J. Med. Sci.*, vol. 36, no. COVID19-S4, pp. S27-S31, 2020. DOI: 10.12669/pjms.36.COVID19-S4.2785, [Online].
- [14] A. Farooq, A. Hakkala, S. Virtanen, and J. Isoaho, "Cybersecurity education and skills: Exploring students' perceptions, preferences and performance in a blended learning initiative," in *IEEE EDUCON*, Porto, Portugal, 2020, pp. 1361-1369. DOI: 10.1109/EDUCON45650.2020.9125213.
- [15] O. Imas, V. Kaminskaya, and A. Sherstneva, "Teaching math through blended learning," 2018. DOI: 10.1109/ICL.2015.7318081.
- [16] R. Olum, L. Atulinda, E. Kigozi, D. R. Nassozi, A. Mulekwa, F. Bongomin, and S. Kiguli, "Medical education and e-learning during COVID-19 pandemic: Awareness, attitudes, preferences, and barriers among undergraduate medicine and nursing students at Makerere University, Uganda," *J. Med. Educ. Curric. Dev.*, vol. 7, 238212052097321, 2020. DOI: 10.1177/2382120520973212, [Online].
- [17] Y. E. V. Hong and L. Gardner, "An evaluation of blended courses: Reflections from undergraduates," 2019. Available: https://aisel.aisnet.org/pacis2019/174?utm_source=aisel.aisnet.org%2Fpacis2019%2F174&utm_medium=PDF&utm_campaign=PDFCoverPage.
- [18] M. T. A. S. K. S. Aditya, and G. K. Jha, "Students' perception and preference for online education in India during COVID -19 pandemic," *SSRN Electron. J.*, vol. 3, no. 1, p. 100101, 2020. DOI: 0.2139/ssrn.3596056, [Online].
- [19] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Thousand Oak, CA: SAGE Publications, 2008, ch. 8, sec. 2, pp. 146.
- [20] T. Chen, L. Peng, X. Yin, J. Rong, J. Yang, and G. Cong, "Analysis of user satisfaction with online education platforms in China during the COVID-19 pandemic," *Healthc.*, vol. 8, no. 3, p. 200, 2020. DOI: 10.3390/healthcare8030200, [Online].
- [21] A. M. Sindiani, N. Obeidat, E. Alshdaifat, L. Elsalem, M. M. Alwani, H. Rawashdeh, A. S. Fares, T. Alalawne, and L. I. Tawalbeh, "Distance education during the COVID-19 outbreak: A cross-sectional study among medical students in North of Jordan," *Ann. Med. Surg.*, vol. 59, no. August, pp. 186-194, 2020. DOI: 10.1016/j.amsu.2020.09.036, [Online].
- [22] F. Brühlmann, S. Petralito, L. F. Aeschbach, and K. Opwis, "The quality of data collected online: An investigation of careless responding in a crowdsourced sample," *Methods in Psychology*, vol. 2, Nov. 2020. DOI: 10.1016/j.metip.2020.100022, [Online].
- [23] B. G. Glaser, "The constant comparative method of qualitative analysis," *Soc. Probl.*, vol. 12, no. 4, pp. 436-445, 1965. DOI: 10.2307/798843, [Online].
- [24] E. Chung and V. N. Mathew, "Satisfied with online learning amidst COVID-19, but do you intend to continue using it?" *Int. J. Acad. Res. Progress. Educ. Dev.*, vol. 9, no. 4, pp. 67-77, 2020. DOI: 10.6007/ijarped/v9-i4/8177, [Online].
- [25] E. Aboagye, J. A. Yawson, and K. N. Appiah, "COVID-19 and e-learning: The challenges of students in tertiary institutions," *Soc. Educ. Res.*, vol. 2, no. 1, pp. 109-115, 2020. DOI: 10.37256/ser.122020422.
- [26] H. H. R. Azami and R. Ibrahim, "Development and evaluation of massive open online course (MOOC) as a supplementary learning tool: An initial study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 532-537, 2019. DOI: 10.14569/ijacsa.2019.0100773, [Online].
- [27] H. H. Razami and R. Ibrahim, "Investigating the factors that influence the acceptance of MOOC as a supplementary learning tool in higher education," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 3, pp. 522-530, 2020. DOI: 10.5373/JARDCS/V12I3/20201219, [Online].

VG4 Cipher: Digital Image Encryption Standard

Akhil Kaushik¹

Ph.D Scholar, CSE Department
Amity University, Gurugram
India

Dr. Vikas Thada²

Associate Professor, CSE Department
Amity University, Gurugram
India

Abstract—When it comes to providing security to information systems, encryption emerges as an indispensable tool, as it has been used extensively in the past few decades for securing stationary data as well as data in motion. With the rapid data transmission techniques and multimedia options available for data representation, the field of information security has become very significant. The state-of-art cryptographic technique is DNA encryption, which uses biological principles for safeguarding data. The use of Bio-inspired ciphers is becoming the de-facto safety standard, especially for digital images as they are a key source of extracting crucial information. Hence, image encoding becomes of ultimate importance when there is a need to send them via an insecure communication channel. The purpose of this research paper is to present a DNA-inspired cryptosystem that can be employed in the domain of image encryption that provides superior security with enhanced efficiency. The experimental outcomes prove that this novel cryptographic algorithm not only provides better security but also at a reasonable pace.

Keywords—DNA cryptography; cipher; information security; encryption; decryption

I. INTRODUCTION

With the epoch of information explosion, information has become the most crucial asset of any individual, corporate, or government. This vital data may contain the personal record of any person, trade secrets of any business organization, or official documents of any government and hence needs to be kept in a secure place. Besides the physical security needs, there is also a need for safety while this significant data during transmission over the vulnerable interaction channel. Cryptography is the remedial solution for such a situation which keeps the information correct and intact between sender and recipient by making the data in a mangled form. Cryptography depends on two things: encryption algorithm and encoding key. The key may be a shared secret key or may form a public-private key pair depending upon the nature of the encryption algorithm. Without the right encryption steps and correct key, the unveiling of secret data can be a herculean task for the adversaries. Thus, cryptography provides impenetrable data which will be meaningless for the eavesdropper [1]. The data is growing immensely as 'big data' and it can now be represented in various forms like text, image, sound, animation, video, etc. This extended volume and multimedia forms of data require modern crypto solutions that can provide information security from malevolent adversaries and uncover the actual information only to the intended recipient. There are a plethora of options available to provide robust security and not all options are suited for all sorts of media. Some ciphers work best on textual data but may perform poorly on video

data, although some security algorithms may perform brilliantly on all sorts of media. Every media form has its peculiar attributes that cause the deviation in the encryption process and outcomes [2].

The most prevalent form of media besides textual data is digital images as they are used widely for information storage and broadcast due to their enhanced data-carrying capacity. The images may contain classified data such as military maps, government documents, healthcare images. The security mechanisms for images can be done in two ways: steganography and cryptography. Image Steganography can be defined as the myriad ways of concealing confidential information in the images, while Image Cryptography alters the image's data to a garbled form making it irrelevant rather than hiding it. Earlier, both of these techniques were used individually, but nowadays they can be combined to provide an even stronger sanctuary solution. Image cryptography is primarily done in two ways: frequency-domain techniques and spatial-domain approaches. The frequency-domain techniques rely on the Fourier transformations, while the spatial-domain approaches simply involve the manipulation of pixels' data in such a way that the real information contained in the pixels of images get altered on the sender's side and gets reverted into the original shape and form on the receiver's side [3]. Apart from the pixel level, encoding can also be done at the bit level, adding more perplexity to the existing encryption systems. There can be innumerable methods of image encryption and the method under consideration in this study makes use of DNA computing.

DNA computing made its way into the modern world when Leonard Adleman experimented with biological data and discovered that biological methods can give productive outcomes while solving baffling computational problems in 1994. Later this phenomenon received greater appreciation and researchers invested their time and money to apply these newly fangled principles into the newer and unexplored domains. DNA cryptography is inspired by the natural process of translation and encoding genetic information in DNA sequences. DNA is the abbreviated form of Deoxyribonucleic Acid that encodes genomic information using enormous sequences of four nucleotide bases $\Sigma = \{A, C, G, T\}$. These chemical bases are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) can be connected into a variety of combinations to pass the genetic information from parents to their children. This information is present in the form of genes which are enormously long DNA distinguishable sequences using complementary pairs of bases and called genes [4], as shown in Fig. 1. There can be multiple ways in which DNA can be utilized in the cryptographic sphere like hiding classified information in long DNA sequences,

generating one-time pads, DNA intensification using Polymerase Chain Reaction (PCR) for using in data encryption, and many more.

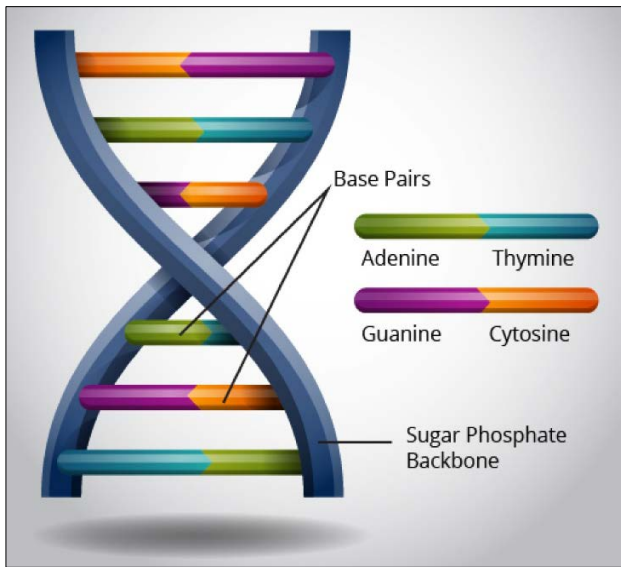


Fig. 1. DNA Structure [4].

The rest of the paper is organized as follows: Section II discusses the background and related work to the research. Section III highlights the system model and adversary model of the proposed research. Section IV details the stepwise proposed methodology for the presented work. Section V presents the simulation outcomes and examines the performance of the proposed cipher. Finally, Section VI entails the conclusion and the direction of future work.

II. RELATED WORK

As the name suggests, DNA cryptography is the amalgamation of biological methods and encryption. As discussed above, DNA computing was introduced in 1994 to solve intricate problems by Leonard Adleman and the concept grew stronger with time. In 1999, Cleland et. al. demonstrated the usage of DNA in steganography by hiding the renowned WWII phrase “June 6 attack; Normandy” in DNA strings [5]. Then, A. Gehani et al. demonstrated the myriad ways to encode data using DNA principles like OTP-based ciphers, Chip-based DNA microarray technology, and DNA steganography also. However, with the period of nearly 20 years, DNA cryptography has achieved new heights and plentiful novel DNA encoding algorithms have been developed, which have been listed below:

One exemplary work was proposed by Suri & Vijay (2017) which mixes a fast chaotic multi-image encryption algorithm with the AES encoding standard for a speedy encoding of multiple images. The concept is unique as it uses Cramer’s rule for decryption of digital images at the receiver’s end [6]. The use of biological methods in image encoding was also suggested by Enayatifar et al. (2017) that first brings the two-dimensional image into a single dimension and later applies permutation and diffusion parallelly for faster encryption. Both operations employ DNA sequences and results show quicker and securer output [7].

Niyat et. al. (2017) recommended the usage of non-uniform Cellular Automata (CA) for image cryptography. This non-uniform CA and hyperchaotic function create a stronger key image that generates a colossal number of random keys. Further usage of chaotic maps to add confusion and later using diffusion principles to create perplex cryptosystem [8]. The encoding of medical images is also a common practice and various studies have been conducted for the same. One such study is done by Akkasaligar & Biradar (2016) which blends DNA with chaotic theory to encode the odd and even pixels of medical images individually. This system is symmetric in nature that maintains integrity and efficiency along with security [9]. Ochani et al. (2017) also worked on the medical images using both steganography and cryptography. The chief technique here is to apply encryption on data and then hiding the patient’s vital encoded data in the cover medical image [10].

Another research done by Wang et. al. (2018) displayed that random numbers can also be used in the image encryption process. The authors have worked on both permutation and diffusion levels to increase the information security, besides including SHA-3 hashing with Chaotic systems to provide ultimate refuse against any unauthorized attacks [11]. Xiuli et. al. (2018) showed the embedding of DNA techniques in encryption. Initially, the image’s color is permuted and concerted into DNA codes. Consequently, DNA is used to produce random numbers that will be used to alter the pre-obtained DNA codes for further diffusion [12]. A similar approach was suggested by Rehman et. al. (2018) for encoding the colored images by combining SHA1 encoding, DNA complementary rules, and chaotic functions. The output obtained exhibited lesser noise and comparatively lower data loss [13].

Two-dimensional Logistic-Sine-Coupling map (2D-LSCM) can also be considered for encryption of colored images. It uses the typical confusion-diffusion structure i.e. transposing the pixels within the image first and then apply diffusion to further modify the pixels. This approach by Hua et. al. (2018) demonstrated better ergodicity than the traditional chaotic systems [14]. Another notable research was carried out by Li et al. (2018) for the encoding of multiple images simultaneously. First of all, Lifting Wavelet Transform (LWT) method is used to produce sparse images and then these scrambled images are XORed to further induce complexity. Finally, the images are compressed to form ghost images that can be traced only through bucket detector arrays [15].

Using DNA in image encryption is also verified by Sun (2018). Primarily, five-dimensional hyperchaotic systems are calculated to generate the chaotic sequences. Then, DNA is combined in several ways like DNA XOR operation, DNA complementary rules, DNA encoding, etc. to enhance the algorithm’s robustness. Besides these superior methods involved, the transposition is also done at two levels: pixel level and binary level [16]. A similar yet different approach was recommended by Zhang et al. (2018) that computed encoding key in two ways i.e. DNA sequencing and logistic chaos mapping. Later, these keys are applied to the plain image using DNA complementary rules to obtain the final ciphered image [17].

Liu et al. (2019) extended the image encryption using DNA to the next level by using 4-D memristive hyper-chaos to generate the chaotic matrices. Then, dynamic DNA is applied on the plain image to produce three matrices, and hence the combination of confusion, diffusion, and encryption creates vigorous cipher [18]. Zhang & Wang (2019) also demonstrated that using a three-dimensional DNA matrix to encode the digital images. First of all, numerous images are combined into a single image and then scramble using chaotic principles. Later on, multiple images are again taken away from the bigger image before applying diffusion using DNA codes and SHA-256 [19].

III. SYSTEM AND ADVERSARY MODEL

A. System Model

Fig. 2 shows the system model considered in this paper, in which there are three major components: Sender's end, Receiver's end, and Genetic Database. The sender's end consists of the user and the machine to generate and encode the message that needs to be sent over to the other end. The message has to be sent over the insecure public communication channel; hence the encoding must be done to keep it intact and away from prying eyes. The receiver's end is quite similar to the sender's end that encompasses the intended recipient and his/ her machine that not only accepts the transmitted message but also decodes it to obtain the original plaintext. However, the process of decryption is somewhat different from encryption depending upon whether it is symmetric or asymmetric cryptography. The third significant component of the model is the Genetic database, which is responsible to create, maintain and securely send the encryption key to the users. There are ample genetic databases that already contain the extremely long DNA sequences stored in the electronic form. The electronic form of DNA strings makes it easier enough for the administrators to store, maintain and give access to the authenticated users. It also eases the user to access, manipulate and regenerate the biological sequences by combining them in several ways. However, genetic databases also play a vivacious role in sharing the secret key that is used by sender and receiver. As the secret key used in the model is deduced from a specific DNA sequence, hence only its

sequence number in the genome database or its URL can be shared with the intended addressee via a safe and trusted channel rather than sending the entire genetic sequence over untrusted communication channels.

B. Adversary Model

In the system model, the real communicate takes place on the public insecure channel; hence there is a stout possibility of attacks to compromise the security. This is due to an imperious factor - an adversary who could try to catch the encoded message and try to either read or manipulate it. An adversary is defined as a computer wizard with malicious intent whose goal is to interrupt or halt the proper functioning of the cryptosystem. In this paper, the following threats are considered:

- **Shared Secret Key Security Threat:** As the communication under consideration is fairly dependent on the shared secret key that is generated using the DNA sequences stored in the genetic databases, hence if the key gets in the wrong hands, the whole system is compromised. Thus, the channel required to send this key must be trusted and secure enough to tackle this threat. Also, the property of backward secrecy should be followed while the generation of session key i.e. knowing one session key should not let the adversary extract other session keys.
- **Privacy Threat:** The message contents need to be kept private although the adversary may feel the message's presence but not its contents. If the adversary can read the message, he may replay the message multiple times or he may be able to modify the message contents or masquerade himself as an authenticated communicating party to gain undue advantage.
- **Physical Security Threat:** The physical security of all the crypto-system components must be of utmost priority as the attacker may try to physically damage or steal the devices or access the system's memory where session keys and messages are stored. Hence, the user credentials' information should be kept integrated and away from the hands of attackers.

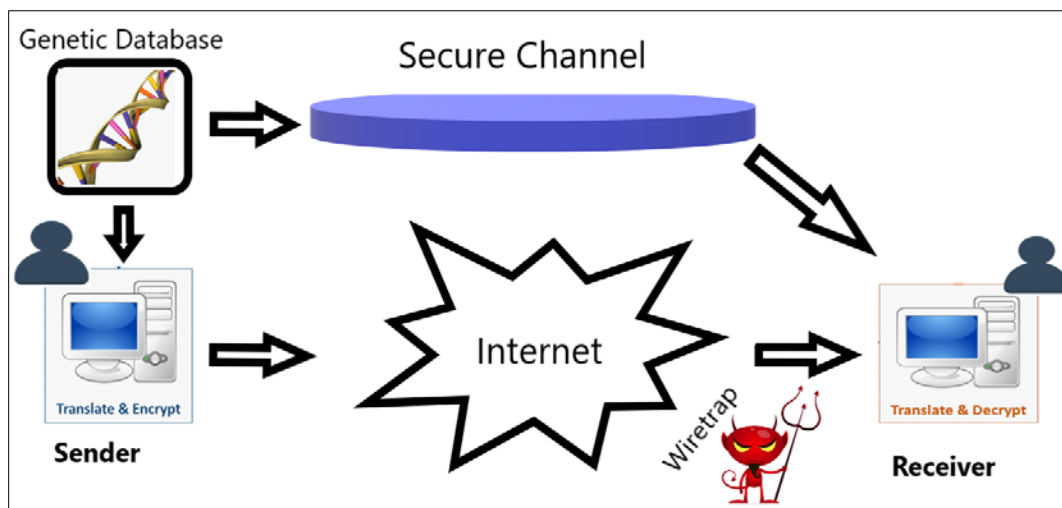


Fig. 2. System Model.

IV. PROPOSED SYSTEM

This section proposes the novel encoding system – VG4 cipher which is based on the existing VG1 cryptographic algorithm. VG1 cipher is a DNA Indexing cryptosystem that is a homophonic substitution cipher for textual data [20]. Several modifications have been done on the VG1 cipher for making it more efficient and secure to apply to digital images.

Image encryption is quite a tedious process with the inclusion of two steps: confusion and diffusion. Confusion refers to the permutation or transpositions of image components (blocks or pixels) so that the correlation among the pixels and the obvious redundancy gets altered. Another vital point is that the confusion step must be reversible to obtain the original image [21].

The second major operation is Diffusion which is stated as the process of altering the value of image components especially the value of pixels to make the encoding robust against differential and noise attacks. The diffusion procedure is done at two levels: block-level and pixel level. The encryption process in detail is described as follows:

- Step 1: The input is a color image of M & N dimensions, where M and N represent the width and height of the image.
- Step 2: The confusion step is done here at block level i.e. grouping of multiple pixels. The block size should be kept an even number and multiple of 8 for ease of operation. Here, the image is divided into 64 blocks and these 64 blocks are shuffled randomly.
- Step 3: There are two encoding keys used in the VG4 cipher. The first key is the ‘Primary Key’ which is deduced from the plethora of DNA sequences available from the genomic databases like GenBank, NCBI, DDBJ, EMBL-Bank, etc. Out of these ample DNA sequences, one peculiar DNA sequence is selected. In this particular DNA sequence, the position of every possible 4-DNA character combination is recorded in a separate dictionary. This dictionary contains position values of a specific DNA byte order (Ex: AATG) in the selected DNA sequence. This dictionary will form a homophonic substitution encryption cipher that works at the pixel level diffusion. An example of such a dictionary is demonstrated in Table I. The subsequent key is the ‘Secondary Session Key’ which is calculated prelude to applying diffusion at the block level. A set of session keys is produced (one for each block) by picking three random symbols from a set of pre-defined characters, $S = \{A-Z, a-z, 0-9, \#, @, !, \$, \%, \wedge, \&, *, \text{etc.}\}$ The secondary session key for each block thus consists of three characters, which are then changed to their binary equivalent.
- Step 4: Prelude to apply diffusion at the block level, a set of session keys is produced (one for each block) by picking three random symbols from a set of pre-defined characters, $S = \{A-Z, a-z, 0-9, \#, @, !, \$, \%, \wedge, \&, *, \text{etc.}\}$ The secondary session key for each block

thus consists of three characters, which are then changed to their binary equivalent. For every block, its combined RGB value is extracted and changed into the binary form so that the corresponding secondary session key can be applied to it. This process is repeated for every block in the image.

- Step 5: After applying diffusion at the block level, the next step is to apply the same at the pixel level. A pixel’s RGB color code in decimal form is acquired using the NumPy library of Python. Using the VG1 encryption process, this decimal value is converted to the binary representation depending on the occurrence and frequency of the decimal values.
- Step 6: The binary data obtained from the previous step is then retransformed using shift-right or shift-left and then coded according to DNA rules (00-c, 01-a, 10-t, 11-g). This DNA encoding output converts all the binary data to DNA form.
- Step 7: The DNA homophonic substitution cipher is then applied on the output of step 6 and hence the 4-letter DNA codes are transformed into decimal numbers (index values in the given DNA sequence). Consequently, decimal values are changed into binary form and finally converted into a ciphered image. The whole encryption procedure is displayed in Fig. 3.

The proposed cryptographic algorithm is primarily based on symmetric encryption and henceforth the decryption procedure of the VG4 cipher is exactly opposite to the encoding process. The genomic sequences from which the primary key is crafted are shared through the reliable secure channel to the recipient. Similarly, the set of secondary session keys are also shared over to the other end. After the generation of encoding keys and receiving the ciphertext by the intended recipient, the next step is to convert the ciphertext into the DNA codes and then to binary form which is rotated into the reverse direction. Subsequently, the data is changed to decimal equivalent and then into pixel’s RGB contents and the original image is conclusively recovered.

TABLE I. EXAMPLE OF KEY INDEXING

DNA Combination	Position Index in the DNA Sequence
GGTA	58, 80, 249, 619, 645, 671, 896, 1197, 1605, 2766, 2958, 2972
AGAG	130, 161, 242, 453, 1011, 1442, 1458, 1512, 1997, 2295, 2789
CAAG	27, 458, 611, 656, 924, 1059, 1332, 1518, 1521, 1539, 1584, 1647, 1695, 1698, 1734, 1767, 1779, 1885, 1933, 2166, 2225, 2365, 2401, 2625, 2700, 2754
AACT	271, 746, 1062, 1188, 1250, 1259, 1409, 1466, 1470, 1491, 1581, 1616, 1701, 1882, 1984, 2095, 2118, 2151, 2198, 2382, 2622, 2655, 2684
AAGG	10, 246, 366, 666, 1182, 1375, 1461, 1448, 1527, 1590, 1593, 1955, 2238, 2338, 2606, 2812, 2864
GGTG	521, 1754, 1877, 1992, 2442, 2531, 2618, 2675

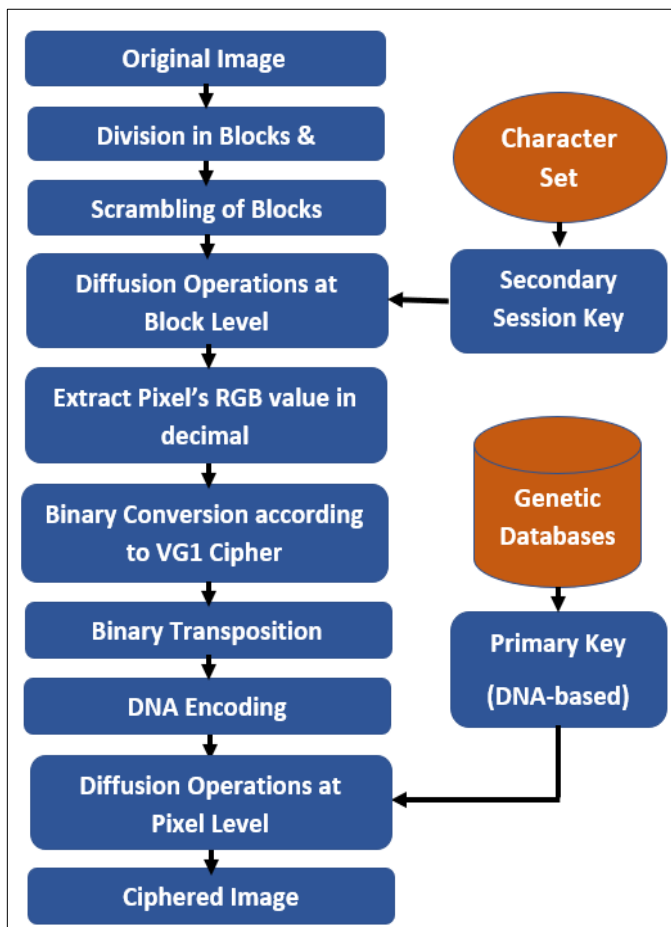


Fig. 3. Basic Block Diagram of VG4 Encryption Cipher.

V. SIMULATION RESULTS AND PERFORMANCE ANALYSIS

The proposed cryptographic algorithm can be implemented in any language that supports Unicode. It is implemented in Python Programming language using Google Colab IDE. It is also implemented on Jupyter Notebook on Intel Core i5 – 10th generation processor HP machine with 8GB RAM. Some crucial evaluation standards that focus on image encryption security are discussed and listed as follow:

A. Key Space Analysis

There is a total of nearly 420 billion DNA sequences available over genomic databases like EMBL-Bank, GenBank, NCBI, etc. Thus, first of all, the user needs to identify which genomic databases are exactly used for determining the primary key, otherwise, he will struggle for a lifetime to unearth it. Even if an attacker discovers DNA string is taken from NCBI, even then using the hit-&-trial method, he/she has to try 4163,000,000 combinations because there are 163 million nucleotide bases in NCBI and there are 4 bases -A, C, G, and T [22]. Thus, the probability of deciding accurate DNA sequence is $\frac{1}{163000000}$. Additional chaos will come into play if a longer biological series is chosen and out of this prolonged sequence only a fraction is extracted for spawning primary key. Hence, the huge keyspace for primary keys makes conventional attacks nearly impossible. Imperative consideration here is that the receiver only needs the correct

DNA number for reproducing the primary key, hence the whole DNA sequence need not be shared over the internet, rather only series numbers can be communicated through secure telephone or any other system.

B. Correlation

The adjacent pixels in any image are correlated to each other and measuring this correlation is of extreme importance when it comes to image cryptography. The input image usually has a high correlation between pixels, while the correlation in the ciphered image is desired to be as low as possible [23]. As depicted in Fig. 4(a), a positive correlation exists between pixels before encryption. However, the results obtained after encoding exhibit that the ciphered images have nearly 0 value of correlation, as shown below in Fig. 4(b). It means that the proposed cipher is successful in weakening the bond between adjacent pixels and makes it harder for the attacker during cryptanalysis.

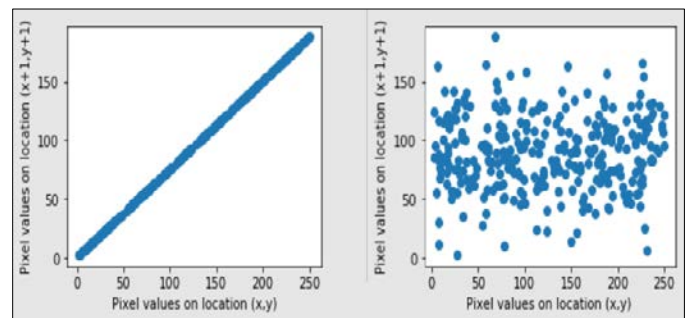


Fig. 4. Correlation between Adjacent Pixels in the Plain Image (a) and Ciphered Image (b) for a 512*512 Size Lena Image.

C. Histogram

The histogram analysis of the image cipher will demonstrate the pixel value distribution and it is desired to be uniform across the whole image. If the variance in histogram decreases in the enciphered image as compared to the plain image, then the cryptosystem is assumed to be fruitful [23]. As clearly observed from Fig. 5, the histogram of the plain image shows non-uniform dissemination, while the encoded image histogram has a uniform distribution pattern.

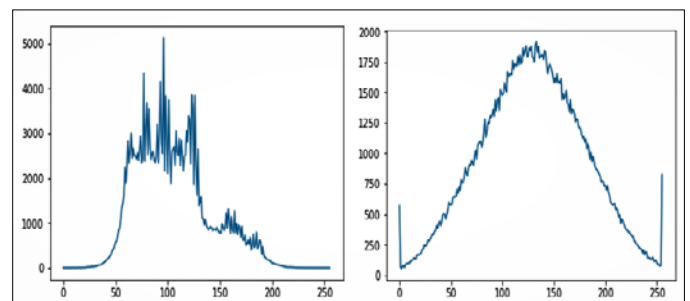


Fig. 5. Histogram of Plain Image (a) and Ciphered Image (b) for a 512*512 size Lena Image.

D. Analysis for Differential Attacks

Differential attacks are based on the idea of tracing the relationship between the original image and image obtained after encryption. Thus, the attacker tries to encode an image to obtain the ciphered image and then make some alterations in

the plain image to observe the subsequent changes in the newly attained encoded image. To measure these changes, there are two quantitative measures: The Number of Pixel Changing Rate(NPCR) and the Unified Average Changing Intensity(UACI). The NPCR for an image of dimension (W * H) is calculated by the following equation [24]:

$$NPCR = \frac{\sum_{i,j} D(i,j)}{W*H} * 100 \quad (1)$$

Where W is Width, H is Height, the value of i is between 1 and W, and the value of j lies within 1 and H. D(i, j) represents the difference between both images. The value of D(i, j) is 0 if the plain and encoded counterparts are the same, otherwise, the value equals 1. The lower bound of NPCR is 0% and the upper bound for NPCR is 100%. The NPCR calculated for VG4 cipher is estimated approximately at 97.65%, which means it is proximate to 100%, hence providing determined robustness.

Another factor UACI is computed as the following [24]:

$$UACI = \frac{1}{W*H} \left[\sum_{i,j} \frac{|C1(i,j) - C2(i,j)|}{255} \right] * 100 \quad (2)$$

Where C1(i,j) and C2(i,j) are the encoded images of plain images with a one-pixel difference. The observed value of UACI is 39.42% which simply means that the proposed cryptographic algorithm is quite sensitive to the minor changes in the plain image and provides substantial security against the differential attacks.

E. Analysis for Noise

The communication channels always contain some kind of noise, which will affect the enciphered image. There can be numerous kinds of noise and quantitative measures to check the effects of these noise on the image quality. The first one in this series is the Mean Square Error (MSE) which is computed as [25]:

$$MSE = \frac{1}{W*H} \sum_{i=1}^W \sum_{j=1}^H [I1(i,j) - I2(i,j)]^2 \quad (3)$$

Where W and H signify the Width and Height of the image respectively. I1(I,j) and I2(i,j) denote the plain image and encoded image correspondingly. MSE value for VG4 cipher is nearly equal to 0.00521.

Depending upon the MSE value, Peak Signal to Noise Ratio (PSNR) can also be computed which is defined as the ratio by which the decrypted image is affected by the noise. Mathematically, it can be defined as [25]:

$$PSNR = 10 \log \frac{(2^n - 1)^2}{MSE} \quad (4)$$

The PSNR value is measured for various noises like Salt & Pepper Noise (SPN), Speckle Noise (SN), and Gaussian Noise (GN) in the units – decibels. The PSNR value for 512*512 ‘lena.jpg’ for noise type SPN with parameters 0.001 is around 39.43. This result indicates detectible recovery from salt & pepper noise of up to 60%.

F. Timing Analysis

The performance analysis of any cryptographic algorithm depends upon security and speed. The encryption and decryption timings of the VG4 cipher are listed in the following Table II:

TABLE II. EXECUTION TIMINGS OF VG4 CIPHER

Image Size	Encryption Timings (Secs)	Decryption Timings (Secs)	Total Execution Timings (Secs)
512*512	40.3	37.6	77.9
700*400	53.2	44.7	97.9
1920*1080	210.6	199.5	410.1

The execution timings of VG4 cipher are compared with other state-of-art encoding algorithms in the following Table III:

TABLE III. COMPARISON OF EXECUTION TIMINGS OF VG4 CIPHER WITH EXISTING STANDARDS

Algorithm	Encryption (secs)	Decryption (secs)
Color image DNA encryption using NCA map-based CML and one-time keys [26]	1300.5	1300.7
Multiple-image encryption via lifting wavelet transform and XOR operation based on compressive ghost imaging scheme [15]	85.2	85.1
An AES-CHAOS-based hybrid approach to encrypt multiple images[6]	407.9	408.2
Multiple-image encryption using genetic algorithm [27]	149.7	144.5
Multiple-image encryption algorithm based on DNA encoding and chaotic system [19]	43	41.3
2D logistic-sine-coupling map for image encryption [14]	9.4	10.2
VG4 Cipher (Proposed)	40.3	37.6

VI. CONCLUSION

In the world of cybercrimes and online scams, a new defense is direly needed to guard the vital information, and for the same, the DNA cryptography approach reinforces the trust back into authenticated users. This paper presents a novel image encryption standard that uses biological principles in addition to traditional cryptography. First of all, the proposed algorithm uses two keys: primary key and secondary key. The primary key is deduced from a digital DNA sequence of long length and provides utmost security against brute-force attacks. The secondary key is also changed per session to enhance robustness. Both diffusion and confusion operations are applied to the image. Confusion or permutation of blocks are applied beforehand and then diffusion is applied at two levels. The secondary key works on a block of pixels and the primary key ensures encoding at the pixel level. The analysis for keyspace, histogram, correlation, etc. has been done and they determine the strength of enciphering algorithm. Also, the analysis for differential attacks and noise has been carried out and the desirable values of NPCR, UACI, PSNR, and MSE show better security and higher resistance against multiple attacks. The timing analysis done against the recent image encryption algorithms has been done and the results demonstrate the novel VG4 cipher is comparable to these modern standards both in encryption as well as decryption timings. Correspondingly, the execution (both encoding and decoding) timings are increasingly linearly with the increasing

size of images, which proves the computational complexity is linear.

In the future, the work can be manifold like chaotic functions like the Chen system or 3D logistic map can be introduced for confusion or diffusion. Another imperative future work could be improving the efficiency of the proposed cipher both in terms of security as well as execution timings.

REFERENCES

- [1] M. E. Saleh, A. A. Aly, & F. A. Omara, Data security using cryptography and steganography techniques (2016).
- [2] M. Jia, Y. Zhou, M. Shi, & B. Hariharan, A deep-learning-based fashion attributes detection model. arXiv preprint arXiv:1810.10148, (2018).
- [3] Z. Hua, Y. Zhou, & H. Huang, Cosine-transform-based chaotic system for image encryption. *Information Sciences*, 480, (2019) 403-419.
- [4] An online article "What is Chemical Structure of DNA" available at <https://empoweryourknowledgeandhappytrivia.wordpress.com/2017/03/29/what-is-the-chemical-composition-of-dna/> (2017).
- [5] A. Y. Niyat, & M. H. Moattar, Color image encryption based on hybrid chaotic system and DNA sequences. *Multimedia Tools and Applications*, 79(1), (2020) 1497-1518.
- [6] S. Suri & R. Vijay, An AES-CHAOS-based hybrid approach to encrypt multiple images, *Recent Developments in Intelligent Computing, Communication and Devices*, Springer, Singapore (2017) 37-43.
- [7] R. Enayatifara, A. H. Abdullah, I.F. Isnin, A. Altameem, & M. Leed, Image encryption using a synchronous permutation-diffusion technique. *Opt Lasers Eng* 90, (2017) 146–154.
- [8] A. Y. Niyat, M. H. Moattar, & M. N. Torshiz, Color image encryption based on hybrid hyper-chaotic system and cellular automata. *Opt Lasers Eng* 90, (2017) 225–237.
- [9] P. T. Akkasaligar, S. Biradar, Secure medical image encryption based on intensity level using chaos theory and DNA cryptography. *International conference on computational intelligence and computing research*, IEEE, Chennai, (2017).
- [10] A. Ochani, D. Jadhav, R. Gulwani, DNA Image encryption using modified symmetric key (MSK). *International conference on inventive computation technologies*, IEEE, Coimbatore, (2017) 1–4.
- [11] X. Wang, S. Wang, Y. Zhang, & C. Luo, A one-time pad color image cryptosystem based on SHA-3 and multiple chaotic systems. *Optics and Lasers in Engineering* 103, (2018) 1–8.
- [12] X. Chai, F. Xianglong, Z. Gan, Y. Lu & Y. Chen, A color image cryptosystem based on dynamic DNA encryption and chaos. *Signal Processing*, (2018).
- [13] A. U. Rehman, X. Liao, R. Ashraf, S. Ullah, & H. Wang, A color image encryption technique using exclusive-OR with DNA complementary rules based on chaos theory and SHA-2. *Optik* 159, (2018) 348–367.
- [14] Z. Hua, F. Jin, B. Xu & H. Huang, 2D logistic-sine-coupling map for image encryption,' *Signal Process.*, 149, (2018) 148-161.
- [15] X. Li, X. Meng, X. Yang, Y. Wang, Y. Yin, X. Peng, W. He, G. Dong, & H. Chen, Multiple-image encryption via lifting wavelet transform and XOR operation based on compressive ghost imaging scheme, *Opt. Lasers Eng.*, 102, (2018) 106-111.
- [16] S. Sun, A novel hyperchaotic image encryption scheme based on DNA encoding, pixel-level scrambling and bit-level scrambling, *IEEE Photon. J.*, 10(2), (2018) Art. no. 7201714.
- [17] T. T. Zhang, S. J. Yan, C. Y. Gu, L. Ren, & K. X. Liao, Research on image encryption based on dna sequence and chaos theory, *Proc. 2nd Int. Conf. Mach. Vis. Inf. Technol. (CMVIT)*, 1004, (2018), 149-154.
- [18] Z. Liu, C. Wu, J. Wang, & Y. Hu, A color image encryption using dynamic DNA and 4-D memristive hyper-chaos, *IEEE Access*, 7, (2019) 78367-78378.
- [19] X. Zhang & X. Wang, Multiple-image encryption algorithm based on DNA encoding and chaotic system, *Multimedia Tools Appl.*, 78(6), (2019), 7841-7869.
- [20] A. Kaushik & V. Thada, VGI Cipher – A DNA Indexing Cipher, *International Journal of Innovative Technology and Exploring Engineering*, 9(3), 2020 221-226.
- [21] Z. Hua, B. Xu, F. Jin, & H. Huang, Image encryption using Josephus problem and filtering diffusion. *IEEE Access*, 7, (2019) 8660-8674.
- [22] E. W. Sayers, R. Agarwala, E. E. Bolton, J. R. Brister, K. Canese, K. Clark, R. Connor, N. Fiorini, K. Funk, T. Hefferon & J. B. Holmes, Database resources of the national center for biotechnology information. *Nucleic acids research*, 47(Database issue), D23, (2019).
- [23] C. Pak, & L. Huang, A new color image encryption using combination of the 1D chaotic map. *Signal Processing*, 138, (2017) 129-137.
- [24] R. Anushiadevi, V. Venkatesh, & R. Amirtharajan, An image mathcrypt-a flawless security via flawed image. *International Conference on Applications and Techniques in Information Security*, Springer, Singapore, (2019) 16-31.
- [25] S. Krivenko, M. Zriakhov, V. Lukin, & B. Vozel. MSE and PSNR prediction for ADCT coder applied to lossy image compression. *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies*, (2018) 613-618.
- [26] X. Wu, K. Wang, X. Wang, H. Kan, and J. Kurths, Color image DNA encryption using NCA map-based CML and one-time keys, *Signal Process.*, 148, (2018), 272-287.
- [27] S. Das, S. Mandal, & N. Ghoshal, Multiple-image encryption using genetic algorithm, *Intelligent Computing and Applications*, 343, (2015).

Formulation of Association Rule Mining (ARM) for an Effective Cyber Attack Attribution in Cyber Threat Intelligence (CTI)

Md Sahrom Abu¹ Aswami Ariffin⁴
Malaysian Computer Emergency Response Team
Cybersecurity Malaysia
Cyberjaya, Selangor DE, Malaysia

Siti Rahayu Selamat² Robiah Yusof³
Faculty of Information Technology and Communication
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

Abstract—In recent year, an adversary has improved their **Tactic, Technique and Procedure (TTPs)** in launching cyberattack that make it less predictable, more persistent, resourceful and better funded. So many organisation has opted to use **Cyber Threat Intelligence (CTI)** in their security posture in attributing cyberattack effectively. However, to fully leverage the massive amount of data in CTI for threat attribution, an organisation needs to spend their focus more on discovering the hidden knowledge behind the voluminous data to produce an effective cyberattack attribution. Hence this paper emphasized on the research of association analysis in CTI process for cyber attack attribution. The aim of this paper is to formulate association ruleset to perform the attribution process in the CTI. The Apriori algorithm is used to formulate association ruleset in association analysis process and is known as the **CTI Association Ruleset (CTI-AR)**. Interestingness measure indicator specially *support (s)*, *confidence (c)* and *lift (l)* are used to measure the practicality, validity and filtering the CTI-AR. The results showed that CTI-AR effectively identify the attributes, relationship between attributes and attribution level group of cyberattack in CTI. This research has a high potential of being expanded into cyber threat hunting process in providing a more proactive cybersecurity environment.

Keywords—*Cyber threat intelligence (CTI); association rule mining; apriori algorithm; attribution; interestingness measures*

I. INTRODUCTION

As the **Tactic, Technique and Procedure (TTPs)** used by an adversary become unpredictable, determined, imaginative, funded, far more coordinated and financially motivated, acquiring useful information from threat information sharing is essential for cyberattack attribution. **Cyber Threat Intelligence (CTI)**, as one of threat information sharing frameworks, has received a lot of media attention in mitigating and reducing cyberattack infection. However, one of the common issues in CTI is the quality of voluminous data from shared information and there is scarce literature in discussing the meaning of quality, basic methods and tools for assessment [1]. A huge volume of data in the CTI consists of raw data without a meaningful relationship between the data. This voluminous data can lead to the ineffectiveness of identifying cyberattack attribution levels due to a lack of useful data from various data sources. Cyberattack attribution process can provide a meaningful relationship between data by identifying the

attribution level and hidden knowledge behind the data to assist organizations in decision making [2]. However, the current cyberattack attribution technique is ineffective in handling the voluminous data in CTI because it relies heavily on the manual process performed by the security analyst and is strictly related to the analyst's knowledge, creating human bias and error-prone [3].

This paper highlight the data mining process in solving the voluminous data issue that can help security analyst to find the relationship between datasets and perform the cyberattack attribution process in CTI. The proposed study was to formulate an association ruleset for cyberattack attribution process in CTI. This ruleset would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level.

The remaining of the paper is organized as follows: Section II presents the research background and related work based on association rules mining in CTI. Section III describes the proposed methodology that includes data collection using CTI feeds, dataset for CTI feeds, association rules mining in CTI framework and formulation of association ruleset using the Apriori algorithm. While Section IV represents the outcome for association ruleset formulation in CTI and evaluate the ruleset generated using interestingness measures. Finally, Section V provides a brief conclusion for this paper.

II. RESEARCH BACKGROUND AND RELATED WORKS

A. Cyber Threat Intelligence (CTI) for Threat Attribution

There has been a lot of studies in the area of data mining to discover its insights in terms of large groups of items or objects in transactional databases, relational databases, or other information repositories using **Association Rule Mining (ARM)** technique. **Association Rule Mining (ARM)** is an important research branch of data mining which has attracted many data mining researchers due to its capability to discover useful and interesting patterns from extensive, noisy, fuzzy and stochastic data. The concept of ARM was introduced by Agrawal and Srikant [4]. In the data mining field, ARM can be utilized as a part of cyberattack attribution process in CTI to discover the hidden knowledge behind raw data. A critical issue for cyberattack attribution in CTI is how to successfully and effectively extract the hidden knowledge from the

voluminous data and feasibly create the association ruleset for cyberattack attribution to assist security analysts in decision making.

Since the introduction of the first concept of ARM by Agrawal et al. [5], a wide variety of efficient ARM algorithms for generating association rules have been proposed over time. Some of the well known and most important algorithms are Apriori, Apriori-TID, SETM, Apriori Hybrid, AIS and Fp-growth [6].

Currently, the most widely used algorithms in ARM is Apriori Algorithm. Agrawal and Srikant developed this algorithm to study customers' purchasing behavior in supermarkets where goods are often purchased together by customers [4]. Besides, the Apriori Algorithm has also been used successfully in many areas of daily life, including energy, recruitment, communication protocol, monitoring and network traffic behavior [7]. Hence, the implementation of the Apriori Algorithm in determining malicious network traffic behavior can help security analysts to study attacker behavior in conducting cyberattack.

Apriori algorithm has been implemented in various fields. Khalili and Sami [8] proposed an industrial intrusion detection approach to mitigate threats to cyber physical systems that utilise sequential patterns extracted by the Apriori algorithm to aid experts in identifying critical states. The study showed Apriori could be employed in the extraction of sequential patterns for industrial process monitoring. A study conducted by Hsiao et al. investigated the use of the Apriori algorithm to track adversaries transitioning through sequences of hosts to launch an attack [9]. Data are retrieved from network packets to determine the host sequence. The Apriori algorithm is proven to be suitable for this study. Meanwhile Liu et al. have utilized Apriori and MS-Apriori algorithm to investigate the relationship of data for network footprint (NFP) which consists of DPI data from ISPs and Crawler data from Web for App usage analysis [7]. The result provides insights for mobile application developers to recommend other applications for their users based on their interest and usage pattern. Adebayo and Abdul Aziz presented a novel knowledge-based database discovery model that utilizes an improvised apriori algorithm with Particle Swarm Optimization (PSO) to classify and detect malicious android application [10]. The usage of several rule detectors can maximize the true positive rate of detecting malicious code, whereas the false positive rate of wrongful detection is minimized. The use of the Apriori algorithm outside the cybersecurity domain has also been explored. It is used for smart health services in a study conducted by Jung, Kim and Chung [11]. The Apriori algorithm was used for a series of patient images acquired through the surveillance technology to generate bio-sequential patient patterns. The bio-sequential patterns are then used to create a basis for a bio-sequential pattern and any deviation from this could result in a possible emergency. The study demonstrated that the Apriori algorithm is used to develop bio-sequential patterns and could be used to extract patterns from the adversary SSH command sequence. Other than that, the Apriori algorithm is also being employed in a study to discover the contributory crash-risk factors of hazardous material (HAZMAT) vehicle-involved crashes on expressways [12]. The findings from this study

indicated that ARM is a feasible technique of data mining that can be used to draw correlations between HAZMAT vehicle-involved accidents and significant crash-risk factors, and has the potential to provide more easy-to-understand findings and applicable lessons for improving the expressways safety.

In this paper, we collect CTI data from current cyberattacks which contained network resources and attackers' behaviour and do association rules analysis using Apriori to generate rules. These rules would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level.

III. METHODOLOGY

In this section, the experimental design to generate the association ruleset in CTI for cyberattack attribution is presented. The input of this experimental design was CTI feeds from OSINT. Data preprocessing technique were used to clean the CTI feeds and produce meaningful data that were used to generate the association ruleset. By conducting this experiment, the association ruleset could be produced to identify the hidden knowledge behind attributes in CTI feeds and identify the attribution level for cyberattack attribution in CTI. The design of experiments is shown in Fig. 1. Fig. 1 illustrates the entire process of association rule mining in CTI framework that consists i) Preprocessing network traffic data, ii) Generating logical rules using Apriori algorithm and iii) Apply the generated rule to facilitate cyber attack attribution. The Apriori Algorithm can discover groups of items occurring frequently together in lots of transactions and such groups of items are called frequent itemsets. The association rule generated from this process is measured using *support*, *confidence*, and *lift*. Given a set of transaction, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) respectively.

To conduct Apriori algorithm on our dataset, we used R to process the filtered data and visualize the result. R is a language and environment for statistical computing, data mining and graphics.

A. Data Collection for CTI Feeds

Data collection for this paper is limited to CTI feeds from OSINT that related to network intrusion activities. For this paper, OSINT CTI feeds from Shadowserver, Lebahnet and MITRE as shown in Fig. 2 has been chosen because it can provide various types of useful information and Indicators of Compromise (IoC) for cyberattack attribution [13]–[15]. The focus of this research was to gather CTI data comprising network resources and attacker behaviour from existing cyberattack.

Fig. 2 shows data collection process for CTI feeds. An API from each CTI feeds was used to collect the data, respectively. Thus, a scraper was used to collect popular network resources such as the domain of search engines or government website, IP address of common DNS server and MD5 hash value of notorious malware from CTI feeds. The examples of attributes collected from each CTI feeds are listed in Table I.

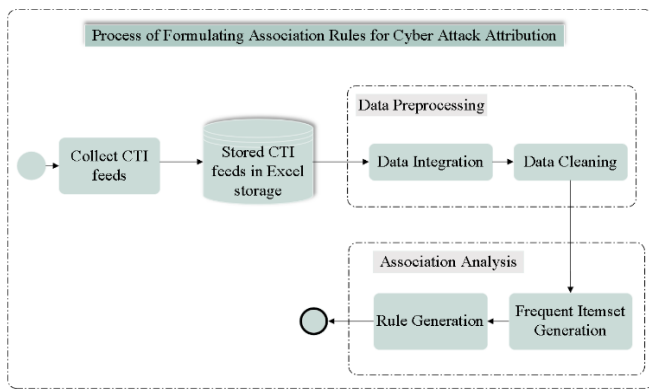


Fig. 1. Experimental Design.

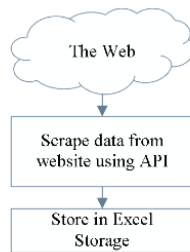


Fig. 2. Data Collection Process.

Shadowserver security feeds provided information about all the infected machines, drones, and zombies that were captured from the monitoring of IRC Command and Controls, capturing IP connections to HTTP botnets, or the IPs of spam relays. Lebahnet security feeds provided valuable supporting information such as network trends and malicious activities that were captured using a collection of distributed honeypots. Both security feeds could provide basic indicators of compromise such as IP address, domain name, URLs, hash value, malware infection type and geolocation. In contrast to Shadowserver and Lebahnet, MITRE knowledgebase was about high-level IOC that related to the behaviour of cybercriminals. MITRE datasets contained various tactics, techniques, software or tools and attackers groups that involved different stages of a cyberattack when infiltrating the network and exfiltrating data. The combination of basic IOC from Shadowserver and Lebahnet and attackers behavior from MITRE knowledgebased was essential in identifying the attribution level for cyberattack attribution in CTI.

B. Dataset for CTI Feeds

The domain of this research was limited to the cyber threat intelligence that related to network intrusion activities and the datasets limited to CTI feeds from OSINT. An API from each CTI feeds was used to collect the data, respectively. Thus, a scraper was used to collect popular network resources such as the domain of search engines or government website, IP address of common DNS server and MD5 hash value of notorious malware from CTI feeds. The CTI feeds covered the top 3 highest infections from 2018 until 2019 in order to be considered relevant cyberattack effort [16]. The summary of each dataset (DS) is depicted in Table II.

TABLE I. THE DETAIL FIELDS IN CTI FEED

(√ = attribute available, x = attribute does not available)
SS=Shadowserver, L=Lebahnet, M=MITRE

Attribute	Data Source			Description
	SS	L	M	
timestamp	√	√	x	the date and time attack captured by the sensor
hashes	x	√	x	malicious file hashes reported in the threat report associated with a particular network resource.
domains	√	x	x	malicious domains reported in the threat report associated with a particular IP resource.
subdomains	√		x	sub-domains reported in a domain resource.
av scans	x	√	x	anti-virus detections reported in the threat report for a network resource or file hash.
source IPs address	√	√	x	malicious IPs used in the attack among the IPs reported in the threat report associated with a particular network resource.
source port number	√	√	x	source of the attacker port number
destination IP address	√	√	x	source of the compromised host
destination port number	√	√	x	source of the compromised port number
URLs	√	x	x	malicious URLs used in the attack among the URLs reported in the threat report associated with a particular network resource.
GeoIP	√	x	x	country of IP or URL location
infection type	√	√	x	malware name as defined by anti-virus detection
technique	x	x	√	technique related to specific threat actors or threat groups
tactic	x	x	√	tactics related to specific actors or threat groups
software or tools	x	x	√	software or tools tactics related to specific actors or threat groups
group	x	x	√	threat actors or groups of threat actors associated with cyberattack

TABLE II. SUMMARY OF THE DATASET FOR EVALUATION

Dataset	Start Date	End Date	Data Source	Total record
DS1	01/05/2018	31/05/2018	Shadowserver	334848
DS2	01/01/2018	31/01/2018	Lebahnet	498
DS3	01/01/2018	31/01/2018	MITRE	15216
DS4	01/03/2019	30/03/2019	Shadowserver	462885
DS5	01/07/2019	31/07/2019	Lebahnet	46
DS6	01/08/2019	31/08/2019	MITRE	4356
DS7	01/06/2018	30/06/2018	Shadowserver	332874
DS8	01/11/2018	30/11/2018	Lebahnet	406
DS9	01/04/2018	31/04/2018	MITRE	21283
DS10	01/07/2019	31/07/2019	Shadowserver	933665
DS11	01/08/2019	31/08/2019	Lebahnet	46
DS12	01/09/2019	30/09/2019	MITRE	5584

Table II shows four datasets from Shadowserver, four datasets from Lebahnet and four datasets from MITRE were collected in this research. The total datasets is twelve and naming as DS1, DS2, DS3, DS4, DS5, DS6, DS7, DS8, DS9, DS10, DS11, and DS12. DS1 to DS3 used for training purposes and explain in Section III (C). While DS4 to DS12 used for evaluation and validation purposes but only result for DS4 explain in Section IV. The rest of DSs were using the same process, hence, adopting the similar explanation as DS4.

C. Association Rule Mining Algorithm in CTI Framework

After the CTI feeds have been preprocessed for producing clean and useful data, the results will be used for association analysis to formulate an association ruleset. This association ruleset is to facilitate a cyber-attack attribution process in the CTI framework to produce an effective threat attribution. The association ruleset can assist security analysts in identifying the origin of the cyberattack and cyberattack attribution level.

To have a general view on the result generated by using R, we set the minimum support value as 0.001 and the minimum confidence value as 0.5. The overall association ruleset analysis classification in CTI was shown in Table III.

The attribution level was divided into three levels namely Level 1, Level 2 and Level 3 [17]. The attributes in Level 1 consisted of IP address, malware type, hash value and port

number, Level 2 was Geolocation and Level 3 needed further analysis of the attributes from Level 1 and 2 to identify the person or attack campaign used by an attacker to launch the cyberattack. However, if the dataset acquired contained the TTP about attackers' behaviour such as datasets from MITRE, then, the attribution for Level 3 was achievable without further analysis from the association ruleset in Level 1 and 2.

Based on the analysis in Table III, three attribution levels can be used to identify the identity and location of an attacker and it can be correlated to CTI type to ease a decision making in an organization.

Table IV depicts the relationship of attribution level and its attribute with CTI types that are useful for verifying the effectiveness of the proposed cyberattack attribution in CTI. Level 1 and Level 2 are parts of tactical intelligence, and the outputs can help an organization to deal quickly and accurately through threatening indicators and prioritize vulnerabilities patches. Level 3 is part of operational intelligence, and its output can improve the detection rate and prevent future incidents as attacks can be seen in a clear context. The conclusion of output from level 1,2 and 3 are part of strategic intelligence which can drive organizations' decision making in terms of security countermeasures and improved areas through comprehending the current attack trends and financial impact to organizations.

TABLE III. OVERALL ASSOCIATION RULESET CLASSIFICATION

(√ = Attribute found, x = Attribute does not found)														
Attribution Level	List of Attribute	Attribute Type							Number of Ruleset in DS					
		IP	hashvalue	URL	Infection type	GeoIP	Technique	Tactic	Software/Tools	Threat actor/Group	DS1	DS2	DS3	
Level 1	'10.0.0.2', '37a98c6150d2317eb6e0df1516a5b3a4', '445', '8a4e9f688c6d0effd0fa17461352ed3e', 'Gen:Variant.Zusy.238725', '1922', '208.100.26.241', '80', 'lethic'	√	√	√	√						7	37	0	
Level 2	'AM', 'MY', 'US'					√					40	0	0	
Level 3	'AppCert', 'Browser', 'COM', 'Component', 'DLLs', 'Distributed', 'Doppelgänger', 'Driver', 'Execution', 'Extra', 'File', 'Hooking', 'Image', 'Injection', 'LSASS', 'Memory', 'Model', 'Mshta', 'Object', 'Options', 'Process', 'Window', 'and', 'apt33', 'cobalt', 'command-and-control', 'credential-access', 'defense-evasion', 'empire', 'execution', 'group', 'lateral-movement', 'mimikatz', 'persistence', 'privilege-escalation', 'strike'							√	√	√	x	0	0	4

TABLE IV. THE ATTRIBUTION LEVEL AND ATTRIBUTE RELATIONSHIP WITH CTI TYPE

Attribution Level	Attribute	CTI type	
Level 1: Cyberweapon	hash value, IP, domain name, URLs	Tactical	Strategic
Level 2: Geolocation	GeoIP		
Level 3: Person or Organization	TTP that consist of technique, tactic, software/tools, campaign name and threat actor name	Operational	

Based on overall association ruleset analysis classification in Table III and attribution level and attribute relationship with CTI type in Table IV, Attribution Level Group for each ruleset (ALGR) is proposed as shown in Table V.

TABLE V. ATTRIBUTION LEVEL GROUP RULESET

Attribution Level Group Ruleset (ALGR)	Description
ALGR1	This group is to represent any ruleset under attribution level 1
ALGR2	This group is to represent any ruleset under attribution level 2
ALGR3	This group is to represent any ruleset under attribution level 3

By using the association ruleset classification in Table III and the proposed ALGR from Table V, the general association ruleset can be defined as an equation (1).

$$\{LHS.A_n\} \Rightarrow \{RHS.A_n\} = ALGR_n \quad (1)$$

Where, $n=$ represent attribution level, Level 1, Level 2 or Level 3; $LHS.A=$ Attribute from attribution level n from the left-hand side, $RHS.A=$ Attribute from attribution level n from the right-hand side, and ALGR = the attribution level group ruleset. While the ruleset representation from the general equation in (1) can be;

$$\{IP, malware\ type, hash\ value\} \Rightarrow \{geolocation\} = ALGR_n$$

$$\left\{ \begin{array}{l} 195.38.137.100, \\ 7867de13bf22a7f3e3559044053e33e7, \\ gamarue \end{array} \right\} \Rightarrow \{RUS\} = ALGR2$$

In this paper, ALGR, as illustrated in Table V and Equation (1), are used to perform cyberattack attribution in CTI.

D. Formulation of Association Ruleset in CTI

In order to prevent cybersecurity threat from causing a significant impact on business and daily life, an actionable threat intelligence with clean data can help an organization in making a fast decision for cyberattack attribution. Cyberattack attribution is defined as a process to identify the location and identity of attackers involved in cyberattack. It is a demanding task that requires a comprehensive intelligence or context to achieve the attribution levels that are divided into three levels namely (1) Attribution to the specific hosts involved in the attack, (2) Attribution to the primary controlling host, (3) Attribution to the actual human actor and attribution to an organization with the specific intent to attack. These attribution

levels can only be achieved when an effective threat intelligence framework is in place. To achieve an effective threat intelligence framework, an organization needs to think of how to build a framework deemed appropriate, specifically, in gaining the hidden information behind the raw data in CTI to assist security analysts in performing cyberattack attribution. Hence, this research focused on formulating an association ruleset in CTI framework to perform cyberattack attribution in CTI. Fig. 3 illustrates Apriori algorithm technique that was used to formulate the association ruleset from CTI OSINT feeds that were collected through CTI framework.

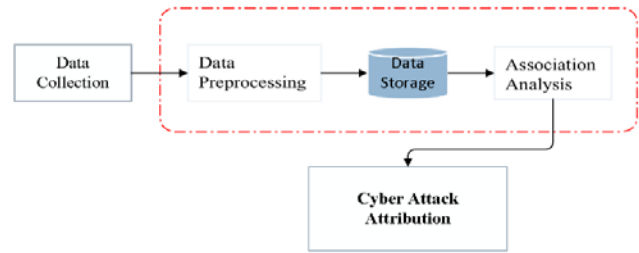


Fig. 3. The Proposed Solution for Cyberattack Attribution in CTI.

Fig. 3 shows the proposed solution to formulate an association ruleset for cyberattack attribution in CTI, which consists of data preprocessing and attribution analysis. The formulation of association ruleset in CTI name as CTI Association Ruleset (CTI-AR) is shown in Table VI.

In Table VI, the purpose and process in CTI framework show that meaningful data that are derived from the preprocessing process are used by the attribution analysis process to identify the attribute and attribution level.

TABLE VI. FORMULATION OF ASSOCIATION RULESET FOR CYBERATTACK ATTRIBUTION PROCESS IN CTI

Criteria	Purpose	Process
Data preprocessing	produce meaningful data, to provide context to raw data	Preprocess raw data
Cyberattack attribution analysis	to identify the threat attribution level, to identify attributes in attribution level	Identify attribute and attribution levels

IV. RESULT

The objective of this section is to present the result of CTI-AR implementation and its effectiveness in performing cyberattack attribution in CTI. This CTI-AR would enable the discovery of hidden knowledge behind the raw data in identifying the attribution level and help security analyst in making a decision for cyber attack attribution. An objective interestingness measure was used to filter and rank the massive amount of association ruleset or CTI-AR generated by Apriori algorithm. This research applied three objective evaluation indicators that were frequently used in Apriori algorithm which were *support* (s), *confidence* (c) and *lift* (l) to measure and determine the interest of ruleset [18]. *Support* reflected the practicality or usefulness of association rules, *confidence* reflected the validity or reliability of association rules and *lift* was to complement previous two evaluation indicators by filtering and removing wrong and meaningless ruleset.

A. Association Rules Analysis for Dataset

The dataset used to mine the frequent itemset was obtained from the ‘Shadowserver security feed’ named “ss_2019_3.csv”. The dataset, dated from 01/05/2018 to 31/05/2018, consisted of malicious network transaction data in Malaysia. It comprised 462885 rows and 35 columns of data, as shown in Fig. 4.

After performing data cleaning by removing incomplete data and filling the missing values, only eight columns of attributes were selected for discovering frequent itemsets as described in Table VII.

Fig. 5 shows a snippet preprocess data for DS4. Apriori algorithm used an iterative level-wise search technique to discover (k + 1)-itemsets from k-itemsets. First, the dataset was scanned to identify all the frequent 1-itemsets by counting each of them and capturing those that satisfy the minimum support threshold. The identification of each frequent itemset required the scanning of the entire dataset until no more frequent k-itemsets was possible to be identified. As for DS4, the minimum support threshold used was 20% or 0.2. Therefore, only the attributes that fulfilled a minimum support count of 0.2 were included in the ruleset generation process.

1	timestamp	dst_ip	port	asn	geo	region	city	hostname	type	infection	url	agent	src_ip	cc_port	cc_asn	cc_geo	
2	2/3/2019 0:00	14.192.212	3468	9534	MY	SELANGOR	PETALING JAYA	http	gamarue	/fnuho	Media/4.1.195.157.15.100					8426 UK	
3	2/3/2019 0:00	118.100.90	2394	4788	MY	MELAKA	MELAKA	http	gamarue	/atomic.php/Media/4.0							
4	2/3/2019 0:00	1.9.247.166	49669	4788	MY	SELANGOR	KUANG	http	gamarue	/forer.php/Media/4.0							
5	2/3/2019 0:00	115.164.204	231	4788	MY	WELAYAH	FUJAJA LAMPUR	http	gamarue	/off.php	/Media/4.0						
6	2/3/2019 0:00	202.188.210	106	50905	MY	SELANGOR	S4AH ALAM	http	gamarue	/forer.php/Media/4.0							
7	2/3/2019 0:00	118.100.70	6432	4788	MY	FULAJU	PERLEBUH DICKENS	http	gamarue	/forer.php/Media/4.0							
8	2/3/2019 0:00	111.121.84	50192	9534	MY	SELANGOR	KLANG	top	mirai							23	
9	2/3/2019 0:00	175.136.22	11080	4788	MY	JOHOR	JOHOR BAHRU	top	mirai								22
10	2/3/2019 0:00	183.171.208	23	35353	MY	SELANGOR	SUBANG JAYA - USJ	12top	mirai								22
11	2/3/2019 0:00	42.188.13	7081	4788	MY	KEDAH	ALOR SETAR	top	mirai								23
12	2/3/2019 0:00	1.92.52.11	28870	4788	MY	SABAH	KOTA KINABALU	top	mirai								29
13	2/3/2019 0:00	115.92.25	52263	4788	MY	SELANGOR	PETALING JAYA - sgp-25-46	top	mirai								2323
14	2/3/2019 0:00	202.188.44	64163	9930	MY	WELAYAH	FUJAJA LAMPUR	top	mirai								22
15	2/3/2019 0:00	42.188.136	3026	4788	MY	SARAWAK	KUCHING	top	mirai								23

Fig. 4. Raw Dataset 4 (DS4).

TABLE VII. DESCRIPTION OF THE ATTRIBUTE FOR DS4

Attribute	Description
timestamp	Timestamp is "DAY MON DD HH:MM:SS YYYY", where DAY is the day of the week, MON is the name of the month, DD is the day of the month, HH:MM:SS is the time of day using a 24-hour clock, and YYYY is the year. The time zone is +0800
dst_ip	Destination IP for infected device
port	Source port of the victim IP connection
geo	The country where the botnet resides
infection	Malware group classification name
src_ip	The IP used by an attacker to manage (C&C) device
src_port	Server-side port for C&C IP
cc_geo	Country of the C&C server

1	timestamp	dst_ip	port	geo	infection	src_ip	cc_port	cc_geo
2	3/2/2019 0:00	14.192.212.162	3468	MY	gamarue	195.157.15.100	22	UK
3	3/2/2019 0:00	118.100.90.229	2394	MY	gamarue	195.38.137.100	22	DE
4	3/2/2019 0:00	1.9.247.166	49669	MY	gamarue	195.38.137.100	22	DE
5	3/2/2019 0:00	115.164.204.231	37304	MY	gamarue	195.38.137.100	22	DE
6	3/2/2019 0:00	202.188.210.106	50905	MY	gamarue	195.38.137.100	22	DE
7	3/2/2019 0:00	118.100.73.140	6412	MY	gamarue	195.38.137.100	22	DE
8	3/2/2019 0:00	121.121.84.222	50192	MY	mirai	195.38.137.100	23	DE
9	3/2/2019 0:00	175.136.226.195	11080	MY	mirai	195.38.137.100	22	DE
10	3/2/2019 0:00	183.171.208.23	35353	MY	mirai	195.38.137.100	22	DE

Fig. 5. Preprocessed Data for Dataset 4 (DS4).

By using the frequent itemset identification process in the Fig. 1, the results of frequent itemsets for DS4 with minimum support count 0.2 were ['195.38.137.100', '22', 'AM', 'DE', 'MY', 'US', 'gamarue']. Then, these frequent itemsets were applied to ruleset generation process as in Fig. 1 to create association ruleset with the predefined minimum confidence (*minconf*) value equal to 50% or 0.5. The value of *minsup*=0.2 and the *minconf*=0.5 were adjusted manually to discover some specific and interesting rules from a large number of random rules [19]. As a result, eighty-one association rules met this threshold configuration. In order to get a realistic overview of the results, the association rules were represented in a scatter plot, as shown in Fig. 6.

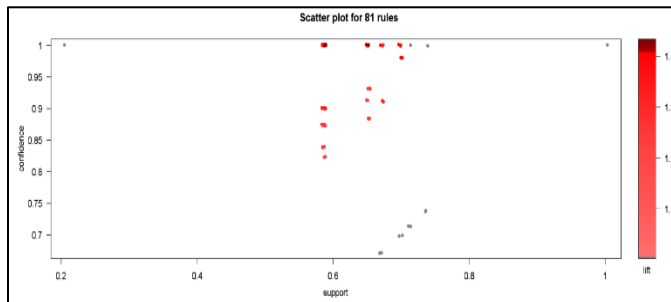


Fig. 6. The illustration of 81 Rules in Scatter plot with Minsup = 0.2 and Minconf = 0.5 for DS4.

Based on Fig. 6, support value represent x-axis and confidence value represent y-axis. For example, the first plot of association rules is located at the coordinate 0.2 for support and 1.0 for confidence. This indicate that the selected plot already meets the threshold for minimum support 0.2 and the threshold for minimum confidence 0.5. To further analyze the relationship between attributes for this association rules, the top five ruleset were selected and presented in Tables VIII, IX, and X based on three IMs; support, confidence and lift. Table VIII shows top five association rules based on support with threshold configured as *minsup* = 0.2 and *minconf* = 0.5.

TABLE VIII. TOP 5 RULES BASED ON SUPPORT MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3, x=Did not meet the requirement to include in attribution level)

No	Ruleset	Support	Attribution Level
R1	{22} ⇒ {MY}	0.74	x
R2	{MY} ⇒ {22}	0.74	x
R3	{DE} ⇒ {MY}	0.71	x
R4	{MY} ⇒ {DE}	0.71	x
R5	{195.38.137.100} ⇒ {DE}	0.7	2

Support could measure the usefulness of association ruleset based on the frequency of itemsets occurring together in the data transaction [20], [21]. The top five rules in Table VIII summed up the combination of rules among port number 22, geolocation MY, DE and IP 195.38.137.100 which indicated that there was a strong association among these four items that frequently occurred together. However, ruleset number R1 to number R4 did not meet the requirement to be included in

attribution level as the implication of antecedents and consequents did not provide meaningful information for decision making. In contrast, the R5 association rule indicated that IP 195.38.137.100 frequently appeared together in the dataset with geolocation DE and provided insight to the security analysts to deduce that the cyberattack possibly originated from this IP and country.

While support measures the usefulness of itemset that is occurring together in data transaction, confidence measure can indicate the strength of association ruleset generated whether it is reliable and valid for decision making [20].

TABLE IX. TOP 5 RULES BASED ON CONFIDENCE MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3, x=Did not meet the requirement to include in attribution level)			
No	Ruleset	Confidence	Attribution Level
R1	{22} ⇒ {MY}	1	x
R2	{DE} ⇒ {MY}	1	x
R3	{195.38.137.100} ⇒ {DE}	1	2
R4	{195.38.137.100} ⇒ {MY}	1	2
R5	{195.38.137.100,DE} ⇒ {MY}	1	2

Table IX presents the top 5 most reliable rules with a threshold for *minsup* = 0.2 and *minconf* = 0.5. The top five rules based on confidence measurement showed that high confidence rules were usually related to port number 22, geolocation MY, DE and IP 195.38.137.100. This ruleset indicated that this IP was used by an attacker to launch a cyberattack and most probably originated from country DE. However, strong association rules are not always effective, some are not what users are interested in, and some are even misleading [21]. For this top five rules only ruleset number R3, R4 and R5 were reliable and were included in attribution level 2.

Support and confidence provided the information about useful rules based on occurrence and reliability of ruleset that occurred in the dataset. Hence, lift measure was needed to complement these two IMs by helping to measure the importance of ruleset that suit the purpose of the research. Table X depicts the top five association rules for lift measure. Three categories were used to interpret the relationship of X / Y in lift measurement. If the lift is equal to 1, then, X and Y are independent. If the lift is higher than 1, then, X and Y are positively correlated. If the lift is lower than 1, then, X and Y are negatively correlated.

Based on Table X, the itemsets of 22, DE, MY, 195.38.137.100 and gamarue respectively had a positive correlation. Thus, this IP was malicious, being infected by gamarue and most probably originated from MY or DE. All ruleset met the requirement to be included in attribution Level 2.

B. Result of Evaluation and Validation for CTI-AR

This evaluation was to determine the capability of the proposed association ruleset for cyberattack attribution process in CTI. However, the number of association ruleset generated

by using the proposed association rule mining could be massive and even tricky for domain specialists to study and summarize the meanings behind the ruleset. Moreover, it was also impractical to sift through a broad set of rules containing noise and irrelevant rules. Hence, the interestingness measure could be used for filtering or ranking association ruleset. This paper only focused on objective interestingness measure using *support*, *confidence* and *lift* to measure the meaningful and reliable association ruleset that were used to guide security analysts in making decisions. The thresholds for minimum support (*minsup*) and minimum confidence (*minconf*) were predefined manually by using trial and error method [7], [19], [22]. The summary of the association rules generated for all the datasets is depicted in Table XI using Apriori Algorithm.

Based on the association ruleset summary, the process of identifying the attributes in attribution level and classifying the ruleset into the respective attribution level group (ALGR) were conducted. Still, not all the generated ruleset met the requirement to be included in the respective ALGR because the ruleset must have at least one attribute from Level 1, Level 2 or Level 3 in both antecedents and consequents.

To further analyzed the findings of evaluation and validation for each association ruleset in Table XII, this paper summarize the ALGR and IM range for DS1 to DS12 in Table XI.

TABLE X. TOP 5 RULES BASED ON LIFT MEASURE WITH MINSUP = 0.2 AND MINCONF=0.5 FOR DS4

(Attribution level: 1=Level 1, 2=Level 2, 3=Level 3 x=Did not meet the requirement to include in attribution level)			
No	Ruleset	Lift	Attribution Level
R1	{22,DE} ⇒ {195.38.137.100}	1.43	2
R2	{22,DE,MY} ⇒ {195.38.137.100}	1.43	2
R3	{DE, gamarue} ⇒ {195.38.137.100}	1.43	2
R4	{22,DE,gamarue} ⇒ {195.38.137.100}	1.43	2
R5	{DE,gamarue,MY} ⇒ {195.38.137.100}	1.43	2

TABLE XI. SUMMARY OF ASSOCIATION RULESET

Dataset	Number of ruleset	Level 1	Level 2	Level 3	N/A
DS1	75	5	40	0	30
DS2	37	0	37	0	0
DS3	4	0	0	4	0
DS4	81	7	40	0	34
DS5	50	45	0	0	5
DS6	12	0	0	12	0
DS7	76	7	40	0	29
DS8	91	89	0	0	2
DS9	14	0	0	14	0
DS10	64	5	31	0	28
DS11	86	84	0	0	2
DS12	17	0	0	17	0

TABLE XII. SUMMARY OF ALGR AND IM RANGE FOR DS1-DS12

(√ = ALGR exist, x = ALGR does not exist) (support=s, confidence=c, lift=l, Attribution Level Group=ALGR)								
Dataset	ALGR			minsup threshold	minconf threshold	Range for IM		
	1	2	3			s	c	l
DS1	√	√	x	0.2	0.5	≥ 0.28	≥ 0.52	≥ 1
DS2	√	x	x			≥ 0.27	≥ 0.52	≥ 1
DS3	x	x	√	0.05		≥ 0.06	≥ 0.53	≥ 2.06
DS4	√	√	x	0.2		≥ 0.21	≥ 0.67	≥ 1
DS5	√	x	x			≥ 0.4	≥ 0.5	≥ 1
DS6	x	x	√	0.05		≥ 0.07	≥ 0.5	≥ 1.84
DS7	√	√	x	0.2		≥ 0.24	≥ 0.52	≥ 0.83
DS8	√	x	x			≥ 0.21	≥ 0.75	≥ 0.96
DS9	x	x	√	0.05		≥ 0.04	≥ 0.5	≥ 1.11
DS10	√	√	x	0.2		≥ 0.37	≥ 0.5	≥ 1
DS11	√	x	x			≥ 0.23	≥ 0.52	≥ 1
DS12	x	x	√	0.05		≥ 0.07	≥ 0.5	≥ 2.86

Table XII shows the range of IM capture from the strongest association ruleset that was generated using the general Equation (1), the threshold used to generate the ruleset and ALGR found in DS4 up to DS12. The value of range for support, confidence and lift in Table XII was used to validate and verify the strong association ruleset to be included in ALGR. Support could measure the usefulness of association ruleset based on the frequency of itemset occurred together in the data transaction. Confidence indicated the strength of association ruleset generated whether it was reliable and valid for decision making. At the same time, lift measure was needed to complement these two IMs by helping to measure the importance of ruleset that suit the purpose of the research, whereby to perform cyberattack attribution process in CTI. Once the list of strong association ruleset was identified and met the threshold for minsup and minconf, this list of association ruleset was included in the respective ALGR based on the presence of the attributes in each association ruleset. The steps to classify the association ruleset into ALGR are explained in the following subsection.

Table XII showed that the ruleset found in this research was effective in performing the cyberattack attribution because it could identify all ALGRs where each ALGR is mapped to different CTI type as discussed in Table IV and Table V. This CTI type was used by an organization for a specific purpose to prevent from cyberattack. For example, ALGR1 and ALGR2 were mapped to tactical intelligence subtype, hence, the outputs from these ALGRs could help an organization to deal with threat indicators and prioritize vulnerabilities patches quickly and accurately. Then, ALGR3 was mapped to operational intelligence and the output from ALGR3 could improve the detection rate and prevent future incidents as attacks could be seen in a clear context. The outputs from ALGR1, ALGR2 and ALGR3 were mapped to strategic intelligence to drive the organization decision making regarding security countermeasure and areas of improvement

from the insights of current attack trends and financial impact to the organization.

The results of the evaluation and validation from the experimental approach are presented in Table XIII. Table XIII illustrates the top 5 association rulesets results from each Interestingness Measure (IM) based on support, confidence and lift measure that filtered and ranked to their respective ALGR. The ALGR grouping could provide hidden information behind the rulesets about attribution level that could help security analysts to perform cyberattack attribution process in CTI.

The association ruleset in Table XIII showed how attributes of LHS implied the attributes of RHS. For example, a ruleset {195.38.137.100,gamarue} ⇒ {22} indicated that an IP address 195.38.137.100 was infected by gamarue and had been used by an attacker to launch an attack using port 22. This ruleset provided the relationship between attribute and guidance to security analysts on the function of the attribute in the cyberattack. This knowledge can help security analysts to plan a mitigation action.

Table XIII also showed how association ruleset were divided into specific ALGR through IM. The grouping of association ruleset into ALGR was based on an attribute that was available in the particular ruleset. Table IV describes the details of attribute in each attribution level. The attributes description for attribution level in Table IV was used as a reference for distinguishing the presence of the attribute from a specific attribution level in each association ruleset. The attribute identification in ruleset could help security analysts to verify what type of attribution achieved from each ruleset. For example, a set of association ruleset in row number four from Table XIII was measured through confidence to prove the reliability of association ruleset provided the information about attribution on IP address, malware type, hash value and port number. The list of attribute found using confidence could be used by a security analyst for further investigation as it is valid and reliable.

Besides, Table XIII also summarized the list of association ruleset into respective ALGR. The classification of ruleset into ALGR was done based on discussion in Table IV and Table V. For example, ruleset classification to ALGR1 was based on the existence of the attribute from Level 1 in the ruleset. This attribute comprised IP address, hash value, malware type, domain name or URLs in the LHS or RHS of the ruleset. As for ALGR2, it required the occurrence of an attribute from attribution Level 1 and Level 2. Geolocation was an attribute of attribution Level 2.

In contrast, the classification of ALGR3 must have attribute from attribution Level 1, 2 and 3 occurred in the ruleset. However, there was also an exception in determining ALGR3, where TTPs alone was sufficient in determining the ruleset as part of ALGR3. It is because TTPs could provide the context to the association ruleset throughout the technique, tactic and procedure used by an attacker to launch the cyberattack.

The results from Table XIII indicated that the formulation of association ruleset from the proposed CTI-AR could help security analysts in making a decision about cyberattack

attribution and the details of the validation result are characterized in Table XIV.

TABLE XIII. RESULTS OF INTERESTINGNESS MEASURE (IM) FOR DS4-DS12

No	Interestingness Measure (IM)	AL GR	Association Ruleset	Attribution achieved
1	Support	1	{195.38.137.100} ⇒ {22} {22} ⇒ {195.38.137.100} {gamarue} ⇒ {195.38.137.100} {195.38.137.100} ⇒ {gamarue} {195.38.137.100, gamarue} ⇒ {22}	IP address, malware type and port number were found
2	Support	2	{195.38.137.100} ⇒ {DE} {DE} ⇒ {195.38.137.100} {195.38.137.100} ⇒ {MY} {MY} ⇒ {195.38.137.100} {195.38.137.100, DE} ⇒ {MY}	IP address and geolocation were found
3	Support	3	{Cloud Service Dashboard} ⇒ {discovery} {discovery} ⇒ {Cloud Service Dashboard} {Cloud Service Discovery} ⇒ {discovery} {discovery} ⇒ {Cloud Service Discovery}	Technique and tactic were found
4	Confidence	1	{210.48.151.111} ⇒ {445} {7867de13bf22a7f3e35590440} ⇒ {Backdoor.Agent.rke} {Backdoor.Agent.rke} ⇒ {7867de13bf22a7f3e35590440} {7867de13bf22a7f3e35590440} ⇒ {210.48.151.111} {7867de13bf22a7f3e35590440} ⇒ {445}	IP address, malware type, hash value and port number were found
5	Confidence	2	{195.38.137.100} ⇒ {DE} {195.38.137.100} ⇒ {MY} {195.38.137.100, DE} ⇒ {MY} {195.38.137.100, MY} ⇒ {DE} {195.38.137.100, 22} ⇒ {DE}	IP address and geolocation were found
6	Confidence	3	{Cloud Service Dashboard} ⇒ {discovery} {Cloud Service Discovery} ⇒ {discovery} {defense – evasion} ⇒ {Application Access Token} {Elevated Execution with Promp ⇒ {privilege – escalation}	Technique and tactic were found
7	Lift	1	{Troj.Spy.Xxp!c} ⇒ {786ab616239814616642ba} {786ab616239814616642ba443} ⇒ {Troj.Spy.Xxp!c} {210.48.151.111, Troj.Spy.Xxp!c} ⇒ {786ab616239814616642ba} {210.48.151.111, 786ab616239814616642ba4438df78a9} ⇒ {Troj.Spy.Xxp!c} {445, Troj.Spy.Xxp!c} ⇒ {786ab616239814616642ba}	IP address, malware type, hash value and port number were found

8	Lift	2	{22, DE} ⇒ {195.38.137.100} {22, DE, MY} ⇒ {195.38.137.100} {DE, gamarue} ⇒ {195.38.137.100} {22, DE, gamarue} ⇒ {195.38.137.100} {DE, gamarue, MY} ⇒ {195.38.137.100}	IP address, malware type, port number and geolocation were found
9	Lift	3	{Elevated Execution with Promp ⇒ {privilege – escalation} {privilege – escalation} ⇒ {Elevated Execution with Pr {Data from Cloud Storage Obj ⇒ {collection} {collection} ⇒ {Data from Cloud Storage C {defense – evasion} ⇒ {Application Access Token}	Technique and tactic were found

TABLE XIV. CHARACTERIZATION OF THE EXPERIMENTAL VALIDATION RESULT

Criteria	Characteristic
Cyberattack attribution analysis	Capable of identifying the relationship of attributes Capable of identifying the attributes in attribution level Capable of identifying the threat attribution level

Therefore, using the characteristics shown in Table XIV, the CTI-AR was validated, as summarized in Table XV.

Table XV indicates the proposed CTI-AR which comprised all characteristics. The proposed CTI-AR was capable of generating the association ruleset from the frequent itemset process, identifying the relationship of attributes among the association ruleset, identifying the threat attribution level for each association ruleset and the attributes in attribution level. Based on the association ruleset and attribution level, the proposed CTI-AR was capable in performing cyberattack attribution process in CTI. These findings were then compared to the findings from the association rule mining (ARM) in existing CTI framework to validate the proposed CTI-AR as discussed in Table XVI.

Table XVI shows the comparison between the association rule mining in existing CTI framework and the proposed CTI-AR in CTI. Based on the characteristics, the ARM in the existing CTI framework is able to identify the attribution level but unable to classify and identify the complete list of attributes that belong to the attribution level. In contrast, the proposed CTI-AR in CTI is more capable in performing the attribution of cyberattacks not only by finding the relationship between the attribute but also providing additional information on the attribution level and attributes at the attribution level.

TABLE XV. SUMMARY OF RESULT VALIDATION OF THE PROPOSED CTI-AR IN CTI

(√ = characteristic exist, x = characteristic does not exist)	
Characteristic	Proposed CTI-AR in CTI
Capable of identifying the relationship of attributes	√
Capable of identifying the attributes in attribution level	√
Capable of identifying the threat attribution level	√

TABLE XVI. COMPARATIVE ANALYSIS WITH EXISTING ARM IN CTI

(√ = characteristic exist, x = characteristic does not exist)		
Characteristics	ARM in existing CTI framework	Proposed CTI-AR in CTI
Capable of identifying the relationship of attributes	√	√
Capable of identifying the attributes in attribution level	X	√
Capable of identifying the threat attribution level	√	√

V. CONCLUSIONS

This paper introduce an approach to overcome voluminous data issue in CTI for cyber attack attribution. The approach consist of data preprocessing, frequent itemset identification and ruleset generation that was used to formulate an association ruleset name as Cyber Threat Intelligence Association Ruleset (CTI-AR). This CTI-AR is used to assist security analyst in discovering the hidden knowledge behind the voluminous data to produce an effective cyberattack attribution in CTI. The results obtained in the experiment demonstrates the CTI-AR is able to discover the hidden knowledge behind the voluminous data in CTI that can help security analyst in performing cyber attack attribution effectively. These abilities are demonstrated through the result obtained using three Interestingness Measures indicators: *support (s)*, *confidence (c)* and *lift (l)*. *Support (s)* reflected the practicality or usefulness of association rules, *Confidence (c)* reflected the validity or reliability of association rules and *Lift (l)* was to complement previous two evaluation indicators by filtering and removing wrong and meaningless ruleset. Based on the result from Interestingness Measures indicators, CTI-AR can effectively help security analyst identify the attributes, relationship between attributes and attribution level group of cyberattack in CTI. This research has a high potential of being expanded into cyber threat hunting process in providing a more proactive cybersecurity environment. For future work, more association rule algorithm and other statistical measures can be implemented to improve association ruleset effectiveness and accuracy in performing cyber attack attribution.

ACKNOWLEDGMENT

This study was kindly supported by The Ministry of Communications and Multimedia (KKMM), Cybersecurity Malaysia and Universiti Teknikal Malaysia Melaka (UTeM).

REFERENCES

[1] C. Sauerwein et al., "Threat Intelligence Sharing Platforms: An Exploratory Study of Software Vendors and Research Perspectives," pp. 837–851, 2017.

[2] S. Qamar, Z. Anwar, M. A. Rahman, E. Al-Shaer, and B.-T. Chu, "Data-driven analytics for cyber-threat intelligence and information sharing," *Comput. Secur.*, vol. 67, pp. 35–58, 2017.

[3] E. C. L. L. W. E. Karafili, "An Argumentation-Based Approach to Assist in the Investigation and Attribution of Cyber-Attacks," *arXiv Comput. Sci.*, 2019.

[4] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules (expanded version). Research Report IBM RJ 9839," *Proc. 20th Intl. Conf. VLDB*, pp. 487–499, 1994.

[5] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," no. December, pp. 207–216, 1993.

[6] P. Prithviraj and R. Porkodi, "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study," *Open J. Comput. Sci. Eng. Surv.*, vol. 3, no. 1, pp. 98–119, 2015.

[7] Y. Liu, K. Yu, X. Wu, Y. Shi, and Y. Tan, "Association rules mining analysis of app usage based on mobile traffic flow data," in *2018 IEEE 3rd International Conference on Big Data Analysis, ICBDA 2018*, 2018, pp. 55–60.

[8] A. Khalili and A. Sami, "SysDetect: A systematic approach to critical state determination for Industrial Intrusion Detection Systems using Apriori algorithm," *J. Process Control*, vol. 32, no. April 2018, pp. 154–160, 2015.

[9] H.-W. Hsiao, H.-M. Sun, and W.-C. Fan, "Detecting stepping-stone intrusion using association rule mining," *Secur. Commun. Networks*, vol. 6, no. March, pp. 1225–1235, Mar. 2013.

[10] O. S. Adebayo and N. Abdul Aziz, "Improved Malware Detection Model with Apriori Association Rule and Particle Swarm Optimization," *Secur. Commun. Networks*, vol. 2019, pp. 1–13, Aug. 2019.

[11] J. C. Kim and K. Chung, "Sequential-index pattern mining for lifecare telecommunication platform," *Cluster Comput.*, vol. 22, no. 4, pp. 1039–1048, 2019.

[12] J. Hong, R. Tamakloe, and D. Park, "Application of association rules mining algorithm for hazardous materials transportation crashes on expressway," *Accid. Anal. Prev.*, vol. 142, no. 3, pp. 105–497, 2020.

[13] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, "Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 755–766.

[14] Z. Zhu and T. Dumitras, "FeatureSmith: Automatically Engineering Features for Malware Detection by Mining the Security Literature," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*, 2016, pp. 767–778.

[15] C. Sabottke, O. Suci, T. Dumitras, C. Sabottke, and T. Dumitras, "Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits," *Proc. 24th USENIX Secur. Symp.*, 2015.

[16] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists," *Secur. Inform.*, vol. 3, no. 1, pp. 1–10, 2014.

[17] Jawwad A. Shamsi, S. Zeadally, F. Sheikh, and A. Flowers, "Attribution in cyberspace: techniques and legal implications," *Secur. Commun. NETWORKS*, 2016.

[18] D. S. S. Mrs. M.Kavitha, "Association Rule Mining using Apriori Algorithm for Extracting Product Sales Patterns in Groceries," *Int. J. Eng. Res. Technol.*, vol. 8, no. 3, pp. 5–8, 2020.

[19] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," *Sci. World J.*, vol. 2014, 2014.

[20] X. Niu and X. Ji, "Evaluation methods for association rules in spatial knowledge base," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. II-4, no. May, pp. 53–58, 2014.

[21] C. Ju, F. Bao, C. Xu, and X. Fu, "A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit," *Discret. Dyn. Nat. Soc.*, vol. 2015, no. 2, pp. 1–10, 2015.

[22] L. Yan, Y. Ke, and W. Xiaofei, "Association Analysis Based on Mobile Traffic," in *2014 4th IEEE International Conference on Network Infrastructure and Digital Content*, 2014.

Integrated Pairwise Testing based Genetic Algorithm for Test Optimization

Baswaraju Swathi¹
Research Scholar
Department of CSE
Jain University, Bengaluru
India

Dr. Harshvardhan Tiwari²
Associate Professor, Centre for Incubation
Innovation, Research and Consultancy (CIIRC)
Jyothy Institute of Technology, Bengaluru
Karnataka, India

Abstract—Generation of Test cases in software testing is an important and a complex activity as it deals with diversified range of inputs. Fundamentally, test case generation is considered to be a multi-objective problem as it aims to cover many targets. Deriving test cases for the Web Applications has become critical to the most of the enterprises. In this paper, a solution for generating test cases for web applications is proposed; the solution uses the System Graph (consisting of links and data dependencies) considering that test cases were based on a combination of input values and data dependencies. Pairwise testing is used to derive the test cases to be executing from entire test cases and then a genetic algorithm is proposed to generate test cases specific to functional testing. The proposed approach was tested through two distinct experiments by measuring the code coverage at every generation and results show that genetic algorithm used increased the fitness value and code coverage. Overall, the results of the paper validate the proposed approach and algorithm, having potential in further construct an automated integrated solution for generating test cases for the entire process.

Keywords—Test case generation; genetic algorithm; multi objective optimization; pairwise testing; test optimization; fitness value

I. INTRODUCTION

The key point regarding the usage of Soft Computing in testing is towards maximizing the quality of software testing and to automate the test generation process. The search issue is all about finding perfect results from a list of adversary results, which is handled by a fitness function to identify results. Software Testing is not just limited to the testing of an application or system but includes checking entirety of a system. Genetic Programming (GP) [1] is a type of Evolutionary Algorithm which is simulated by biological growth to search programs that perform certain user-defined tasks. This programming technique has been successfully applied to many fatigue problems present in software testing such as instinctive design, pattern recognition, and test suit generation [2]. This suit of algorithms helps to automate the generation of basic test paths which includes several problems like data generation, sequence generation, test case derivation, and optimization. Recent studies stated a regeneration genetic algorithm which is proven to be operative and trivial for coverage-oriented software test suit generation. Issues related to software reusability can be resolved by the grouping of soft computing approaches like neural networks through software

testing. Soft computing techniques such as GA are very much suitable for test size and coverage problems. Efforts are taken to develop the finest potential solutions for the automation in test suits and test sequence generation. Problems like test sequence, test data generation in white box testing and functional testing uses GA. In the current era of software development, test automation has a significant function in testing the software in its entirety. Test automation comes with its own challenges which include reusable scripts generation, recompiling the test scripts with modifications for different runs and rapid test development with least amount of development time and effort. Traditional methods such as randomized approaches, goal aligned techniques involve human intervention, development effort, cost. Limited Resources, missing the critical requirements and generation of redundant test cases are the prominent constraints in test case generation. To overcome the mentioned challenges test case generation methods needs enhanced algorithms.

II. RELATED WORK

Test case generation techniques are classified into specification based which uses specification documents to derive the test cases, sketch based commonly work with diagrams such as UML, source code based where in test cases are derived using source code applicable to white box testing[3]. Study suggests test case generation to be a complex problem where in various strategies were proposed for the same. The algorithms GA, GA-NN and MA algorithms were applied in [4] which applies Machine Learning techniques to test generation process. Sketch based test case generation in combination with uml diagrams and state transition diagrams were proposed by [5].[6-8] Test case in combination with soft computing techniques such as Genetic Algorithm, Particle Swarm Optimization, Artificial Bee Colony derived a suitable results. The proposed approach considers test case generation process as a combinatorial optimization and the best feasible solution is in a set of discrete range. Combinatorial solutions were present in the literature, with different approaches: single objective optimization, multiple objective optimization. Test Case is a set of various combinations of input values which run on a scenario to produce the result and later decided accordingly. Hence the Test Case problem is (T, U, M, F), T is set of test instances can be considered as a test set, U is determinate set of solutions from the suite, given an instance x and a feasible solution y m is a measure on y.

Dependency Coverage: Data dependencies and link dependencies [23] drive the extreme amount of test cases where page transits throughout the web application.

IV. STRUCTURE OF GENETIC ALGORITHM

Genetic Algorithm runs with an initial population of genes which are test cases. Fitness function is computed over the population to select set of chromosomes which will participate in the next generation population. Cross over and mutation operators are applied over the selected population to generate diversified range of population. GA stops once the population is either converged or for a specified number of iterations.

A. Genetic Algorithm

GA is a parameter coding technique which usually works on population of solutions and deterministic transitions. Considering the test case generation with respect to multi objective optimization PARETO [24] solutions and multi-criteria decision-aid technique is applied to select the finest solution. PROMETHEE technique [25] of decision is applied such that ranking amongst the individuals. Positive ranking is given as in Eq. 1, which expresses to what extent each alternative outranks all the others.

$$\sum_{b \in A, b \neq a} \pi(a, b) \phi(a) = \frac{1}{n} - 1 \quad (1)$$

B. Genetic Algorithm Encoding

Each chromosome is encoded as a combination of pages and the data flow for each element to other element in a web page. We use a graph data structure to indicate the paths and web pages. The data flow from one element to other likely one page to other page is created. $P1 \rightarrow P2 \rightarrow P3 \dots Pn$. From the Graph below sample genes considered:

$$\begin{aligned} \text{geneA} &= \{0, 1, 5, 8, 5, 9\} \\ \text{geneB} &= \{5, 4, 7, 5, 6\} \\ \text{geneC} &= \{1, 5, 2, 5, 6, 10, 6, 3\} \\ \text{geneD} &= \{0, 1, 5, 3\} \end{aligned}$$

C. Fitness Function and Selection Mechanism

Tournament based selection [26, 27] is preferred over the roulette wheel selection as to lessen the risk of missing test cases. The primary fitness value is derived based on the valuation standard code coverage. If selected set of test cases covers the maximum code coverage are assigned to be highly probable. Secondary fitness value is dependent on the number of data dependencies and link dependencies of the given nodes. Individual gene with fitness f will succeed in the tournament of t individuals picked from the test suite with whole population given as in Eq. 2.

$$P(F) = \text{MAX}(F1, F2, \dots, FN) = X P(F < H)^{S-1} P(F) \quad (2)$$

where $P(F)$ constitutes the probability. S denotes the genes having lower fitness score. The anticipated tournament succeed from a tournament size s is specified as in Eq. 3.

$$s \int f P(f s - 1 p(f)) df \quad (3)$$

A test case is given a higher fitness value depending on the below functions.

Code Coverage of the chromosomes

$nd \rightarrow$ Number of data dependencies,

$nl \rightarrow$ Number of link dependencies,

tnd and tnl are the all-inclusive number of data dependencies and all-inclusive number of link dependencies contributed by the test case.

Test cases corresponding to the genes defined in the above section:

geneA : { 0, 1, 5, 8, 5, 9 }

TC1: $P1 \rightarrow P2 \rightarrow P6 \rightarrow P9 \rightarrow P6 \rightarrow P10$

geneB = { 5, 4, 7, 5, 6 }

TC2: $P6 \rightarrow P5 \rightarrow P8 \rightarrow P6 \rightarrow P7$

geneC = { 1, 5, 2, 5, 6, 10, 6, 3 }

TC3: $P2 \rightarrow P6 \rightarrow P3 \rightarrow P6 \rightarrow P11 \rightarrow P7 \rightarrow P4$

geneD = { 0, 1, 5, 3 }

TC4: $P1 \rightarrow P2 \rightarrow P6 \rightarrow P4$

Cross over: Single point crossover is considered initially to generate new population, if diversified range of population to be generated the other cross over operations can be applied.

TC1: $P1 \rightarrow P2 \rightarrow P6 \rightarrow P9 \rightarrow P6 \rightarrow P10$,

TC2: $P6 \rightarrow P5 \rightarrow P8 \rightarrow P6 \rightarrow P7$

TC3: $P2 \rightarrow P6 \rightarrow P3 \rightarrow P6 \rightarrow P11 \rightarrow P7 \rightarrow P4$,

TC4: $P1 \rightarrow P2 \rightarrow P6 \rightarrow P4$

TC11: $P1 \rightarrow P2 \rightarrow P5 \rightarrow P8 \rightarrow P6 \rightarrow P7$ (TC1&TC2)

TC12: $P2 \rightarrow P6 \rightarrow P9 \rightarrow P6 \rightarrow P10$ (TC1&TC3)

The mutation process [28] is to maximize the chance of complete search space in the algorithm, a predefined mutation probability [29-30] is calculated for each chromosome, and score is arbitrarily engendered to relate the mutation probability to resolve for the mutation process. Sample of the test cases after the crossover operation and mutation operation.

From the above generated test cases:

TC21: $P4 \rightarrow P6 \rightarrow P9 \rightarrow P6 \rightarrow P10$ (TC1 and TC4)

Acceptance: As the mutation and crossover involve certain level of uncertainty, the off springs may or may not be superior to parent chromosomes. Hence fitness needs to be calculated for acceptance.

Stop criteria: for a specified number of maximum generations the GA is executed, based on the fitness and code coverage the GA is stopped.

ALGORITHM - TEST CASE GENERATION

Input:

Program under test
Initial set of paths (Test Cases) from the System Graph for Web application.
Initial set of paths (Test Cases) from the Program Graph for console programs.

Output:

Set of optimized paths (Test Cases),
{P1, P2, P3...Pn}, Code Coverage, Fitness value.

Initialization phase:

Build a DLDG graph for the corresponding program under test.

Generate initial population of genes

{TC1, TC2, TC3...TCn},

Apply Pairwise testing to generate genes

{TCm1, TCm2, TCm3...TCmn}

GA Algorithm:

gen1=1, max_gen

Current_population:

{TCm1, TCm2, TCm3...TCmn}=

{P1→P2→P3..Pi}, (initial set of paths)

While (gen1≤ max_gen)

Begin

for each gene gene_i in Current_ population

{gene₁, gene₂, gene₃...gene_n}

Calculate the fitness_value Fi as specified in Eq.4

$$Fi = \sum_{i=0}^n Ci + TCi((nd + nl)/(tnd + tnl)) \text{ Eq. (4)}$$

Ci is the code coverage of the test suit, nd, nl,tnd,tnl as stated in the fitness and selection mechanism.

for each gene{gene_i}

If (fitness_value is in the range)

Select the gene {gene_i} based on Tournament based selection

Apply crossover operation to generate the new genes

Apply mutation operation to change the gene

Add the above population to the current_ population

End

V. EXPERIMENTS AND EVALUATION

Experiment 1:

Triangle classification problem where in the input is considered for three sides of a triangle and the output details the type of a triangle. SideA, SideB, SideC for the first generation was chosen randomly as specified in Table I, these values were further selected to be part of parent chromosomes and underwent GA operations using the fitness function and pairwise testing described in the above algorithm. Pairwise testing values were obtained using online Pairwise online tool. Code coverage from the second generation was noted and specified in Table II. NUnit coverage tool is used to record the code coverage of the test suit. The tables provide the data obtained as a result of our methodology in Fig. 1.

Fig. 3 illustrates the tests vs coverage in terms of line and branch coverage for the values specified on the horizontal axis.

TABLE I. GENERATION- 1

SideA	SideB	SideC
0	2	1
2	5	1
1	1	5
5	2	1

TABLE II. GENERATION-2

Test case SideA,SideB,SideC	Branch coverage	Line coverage
1,2,1	33.33%	40.90%
1,0,0	33.33%	31.81%
1,5,0	25%	31.81%
1,1,2	33.33%	31.81%
1,1,0	25%	31.81%
1,0,2	33.33%	31.81%
2,1,1	41.66%	40.90%
2,5,2	41.66%	40.90%
2,0,1	33.33%	31.81%
2,1,1	33.33%	31.81%
2,2,0	25%	31.81%
5,5,1	41.66%	40.90%
5,0,1	33.33%	31.81%
5,1,1	33.33%	31.81%
5,1,0	25%	31.81%
5,2,1	50%	45.45%
5,1,2	50%	45.45%
0,0,1	25%	31.81%
0,1,0	25%	31.81%
0,1,1	25%	31.81%
0,2,2	25%	31.81%
0,5,1	25%	31.81%

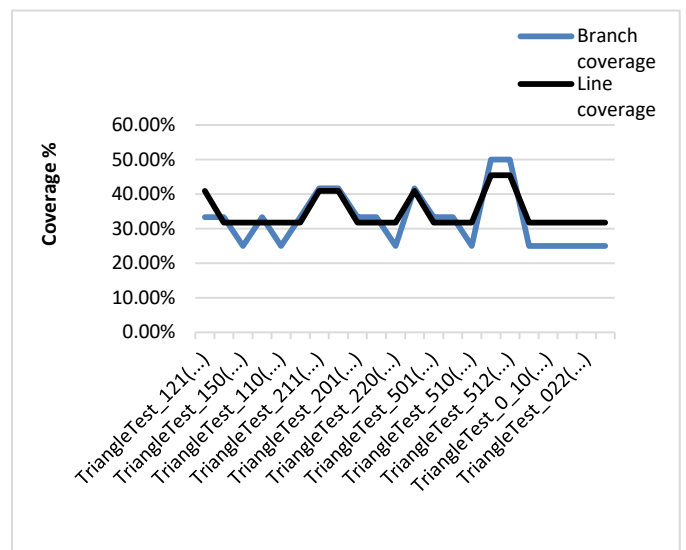


Fig. 3. Tests vs. Coverage.

The sample values after processing and normalized values achieved the below result as shown in Fig. 4.

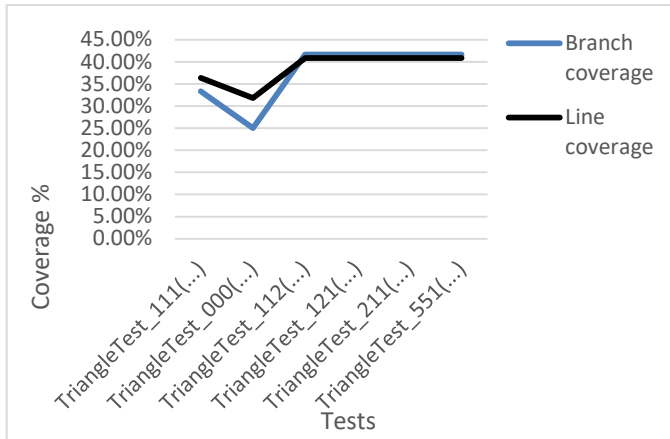


Fig. 4. Tests vs. Coverage.

The results after eight generations achieved a consistent result which achieved 88.20% of code coverage and are depicted in the Fig. 5.

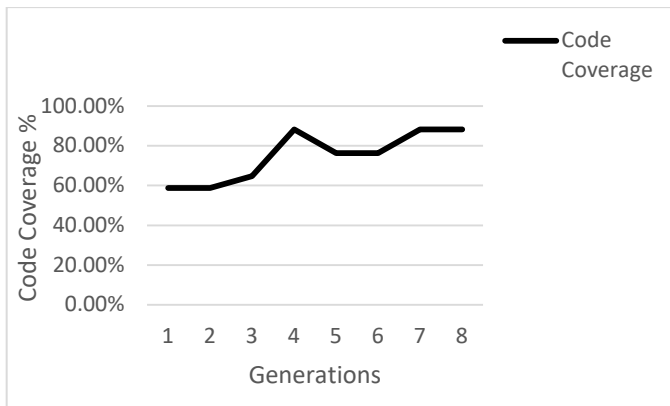


Fig. 5. Generation vs. Code Coverage.

Experiment 2:

The source code for a simple web application was considered for experimental evaluation and random test cases were generated. This was a Web based application as represented in Fig. 2, the automated test cases were captured using selenium IDE. Selenium IDE is basically a record and playback tool, the test cases generated by Selenium IDE are saved and deployed as JUnit, NUnit test cases. The sample test cases were run through NUnit code coverage [31, 32] which achieved the following result over the main modules like performing an insertion and deletion of the records of customers.

Proposed GA Algorithm was then executed on the same set considering few sample test cases from the above document, which achieved the following result. At each iteration the fitness value is generated using the fitness function and the genes are allotted the ranking as per selection criteria discussed previously. The set of genes which are valid and invalid is checked manually which can be automated further. Hence the genes undergo a preprocessing phase for the mentioned.

Considering the above mentioned geneA, geneB, geneC, geneD, Sample of genes generated by GA algorithm for three of the generations are as mentioned below.

Test cases derived for a sample of three generations

Generation 0: Random population considered from the Fig. 3 are:

- [0, 1, 5, 8, 5, 9]
- [5, 4, 7, 5, 6]
- [1, 5, 2, 5, 6, 10, 6, 3]
- [0, 1, 5, 3].

After processing with the selection, crossover and mutation operation the following were the chromosomes generated for second generation.

Generation 1:

- [0, 1, 5, 5, 6, 9]
- [5, 4, 7, 8, 5]
- [0, 1, 5, 8, 5, 9]
- [5, 4, 7, 5, 6]
- [0, 1, 5, 5, 6, 9]
- [5, 4, 7, 8, 5]
- [0, 1, 5, 8, 5, 9]
- [5, 4, 7, 5, 6]
- [0, 1, 5, 3, 6, 10, 6, 3]
- [1, 5, 2, 5]
- [0, 5, 2, 5, 6, 10, 6, 3]
- [1, 1, 5, 3]
- [0, 5, 5, 3, 6, 10, 6, 3]
- [1, 1, 2, 5]
- [0, 5, 5, 5, 6, 10, 6, 3]

Generation 2:

- [0, 5, 5, 5, 6, 10]
- [0, 1, 5, 8, 5, 9, 6, 3]
- [0, 1, 5, 8, 5, 9]
- [0, 5, 5, 5, 6, 10, 6, 3]
- [0, 1, 5, 5, 6, 10]
- [0, 5, 5, 8, 5, 9, 6, 3]
- [0, 1, 5, 8, 5, 9]
- [0, 5, 5, 5, 6, 10, 6, 3]
- [1, 1, 2, 3, 6]
- [5, 4, 7, 5]
- [1, 4, 7, 5, 6]
- [5, 1, 2, 3]
- [1, 4, 2, 3, 6]
- [5, 1, 7, 5]
- [1, 4, 2, 5, 6]
- [5, 1, 7, 3]

The above values after preprocessing where in repeated genes and invalid genes were processed and further reduced. Validity and invalidity of the genes were verified based on the data associated with the genes, for instance the path from P1→P2 is valid based on data which were minimized using pairwise testing. Considering P1 to be a Login page the page transits to other page if P1 {data} is valid, if P1 {data} is not

valid the page transits to other page. Fig. 6 and Fig. 7 depicts the graph Tests vs. coverage, the very first initialization of the test suit is chosen randomly specified with values in the horizontal axis, where the line coverage and branch coverage are proportional, the intermediate tests didn't achieve the coverage but stabilized in due evolution with Genetic Algorithm.

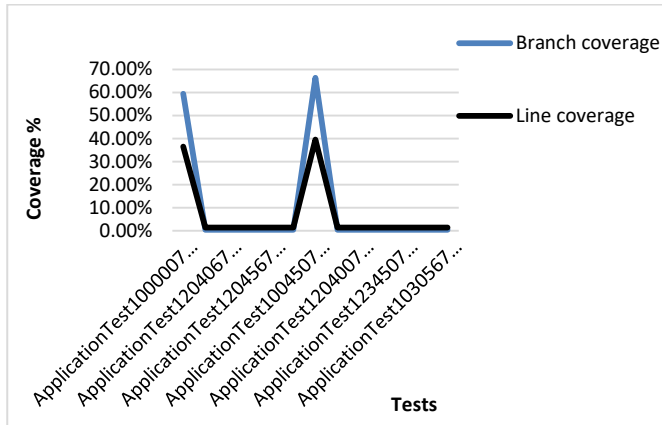


Fig. 6. Tests vs. Coverage for First Generation (Random Test Cases).

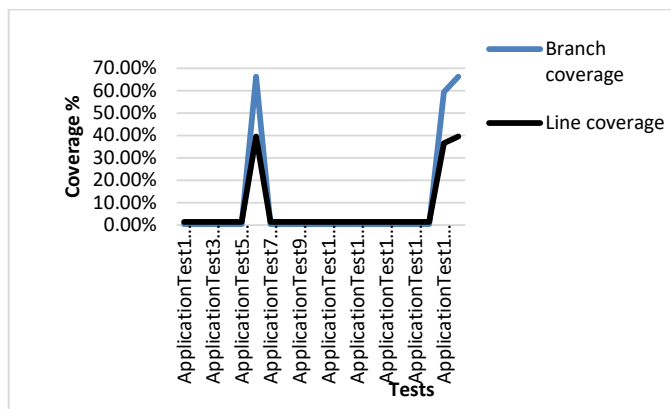


Fig. 7. Tests vs. Coverage for Generation 2.

VI. FUTURE SCOPE

The future work of the proposed work is to evaluate with large scale web applications and console programs. Though the current approach proposes an automated solution, the pairwise integration, validation for each run with respect to the test cases is done manually. The work can be extended with a complete automated integrated solution for generating test cases for the entire process.

VII. CONCLUSION

This paper proposes an automated solution for Test case generation problem by means of Integrated Pairwise Genetic algorithm. A set of optimized test cases after Pairwise testing are considered as initial population for the GA. Considerably less genes were initiated which leads gradually to huge amount of test suites. Code coverage was measured at every generation and based on fitness values the parent genes were selected and then were involved in the generation process. When the evaluation metric code coverage is compared with random

generation of test cases and GA, the results show that GA has considerably increased the fitness value and code coverage. Further our work requires and automated integrated solution for the whole process.

REFERENCES

- [1] Katoch, S., Chauhan, S.S. & Kumar V, "A review on genetic algorithm: past, present, and future," *Multimed Tools Appl* (2020). <https://doi.org/10.1007/s11042-020-10139-6>.
- [2] TianTian and Dunwei Gong. 2016., "Test data generation for path coverage of message-passing parallel programs based on co-evolutionary genetic algorithms," *Automated Software Engg.* 23, 3 (September 2016), 469–500. DOI:<https://doi.org/10.1007/s10515-014-0173-z>.
- [3] NichaKosindrdecha and JirapunDaengdej, 2010, "A Test Case Generation Process and Technique," *Journal of Software Engineering*, 4: 265-287.
- [4] M.R. Keyvanpour, H. Homayouni and HaseinShirazee, 2011, "Automatic Software Test Case Generation," *Journal of Software Engineering*, 5: 91-101.DOI: 10.3923/jse.2011.91.101
- [5] Khurana N, Chillar RS, " Test Case Generation and Optimization using UML Models and Genetic Algorithm," *Procedia Computer Science* [Internet]. 2015;57:996–1004.Available from: <http://dx.doi.org/10.1016/j.procs.2015.07.502>.
- [6] Rijwan Khan, Mohd. Amjad, "Automatic test case generation for unit software testing using genetic algorithm and mutation analysis," *2015IEEE UP Section Conference on Electrical Computer and Electronics (UPCON)*.
- [7] Shveta Parnami, KrishnaSwaroop Sharma, "Empirical Validation of Test Case Generation based on All- Edge Coverage Criteria,"*International Journal of Computer Applications*,September 2015.
- [8] Baswaraju Swathi, Dr.Harshvardhan Tiwari, "Genetic Algorithm Approach to Optimize Test Cases," *International Journal of Engineering Trends and Technology* 68.10(2020):112-116.
- [9] Hulme A, Mclean S, Salmon PM, Thompson J, Lane BR, Nielsen RO, " Computational methods to model complex systems in sports injury research: agent-based modelling (ABM) and systems dynamics (SD) modelling," *Br J Sports Med.* 2019 Dec;53(24):1507-1510. doi: 10.1136/bjsports-2018-100098. Epub 2018 Nov 17. PMID: 30448782.
- [10] Abu Sharkh, M., Shami, A. &Ouda, A, "Optimal and suboptimal resource allocation techniques in cloud computing data centers," *J Cloud Comp* 6, 6 (2017). <https://doi.org/10.1186/s13677-017-0075-2>.
- [11] Islam, M.R., Mahmud, M.R. &Pritom, R.M, "Transportation scheduling optimization by a collaborative strategy in supply chain management with TPL using chemical reaction optimization," *Neural Comput&Applic* 32, 3649–3674 (2020). <https://doi.org/10.1007/s00521-019-04218-5>.
- [12] Ram Krishna Rathore, Kaushal Sharma , Amit Sarda, "An Adaptive Approach for Single Objective Optimization,"*Ram Krishna Rathore et al Int. Journal of Engineering Research and Applications* ,ISSN : 2248-9622, Vol. 4, Issue 2(Version 1), February 2014, pp.737-746.
- [13] AnnibalePanichella , Fitsum MesheshaKifetew , "Automated Test Case Generation as a Many-Objective Optimization Problem with Dynamic Selection of the Targets,"*IEEE Transactions on Software Engineering* (Volume: 44 , Issue: 2 , Feb. 1 2018).
- [14] Libiao Zhang, Xiangli Xu, Chunguang Zhou, Ming Ma, Zhezhou Yu, "An Improved Differential Evolution Algorithm for Optimization Problems,"*Advances in Computer Science, Intelligent System and Environment*.
- [15] HasanUral,KassemSaleh, Alan W Williams, "Test generation based on control and data dependencies within system specifications in SD," *Computer Communications* 23(7):609-627.
- [16] Kamal Z Zamli, "T-Way Strategies and Its Applications for Combinatorial Testing," *International Journal on New Computer Architectures and Their Applications (IJNCAA)*1(2): 459-473The Society of Digital Information and Wireless Communications, 2011 (ISSN: 2220-9085).

- [17] Feng Duan et al, "An Approach to T-Way Test Sequence Generation with Constraints," 2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW).
- [18] Haralambi Haralambiev et al, "Applying source code analysis techniques: A case study for a large mission-critical software system," 2011 IEEE EUROCON - International Conference on Computer as a Tool.
- [19] Rahma Mahmood, Qusay H. Mahmoud, "Evaluation of Static Analysis Tools for Finding Vulnerabilities in Java and C/C++ Source Code," arXiv.org cs, 1805.09040, Cornell University.
- [20] Rahm Mitrabinda Ray, "PSO based test case generation for critical path using improved combined fitness function," Journal , Volume 32, Issue 4, May 2020, Pages 479-490.
- [21] Marko Ivanković et al, "Code Coverage at Google," Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ACM, pp. 955-963.
- [22] Giovanni Grano et AL, "Branch coverage prediction in automated testing," *Journal of Software Evolution and Process*, 08 March 2019.
- [23] Matteo Biagiola et al, "Web Test Dependency Detection," *Coronell University*, arXiv:1905.00357.
- [24] Mark Harman, Kiran Lakhotia, Phil McMinn, "A Multi-Objective Approach To Search-Based Test Data Generation," GECCO'07, July 7-11, 2007.
- [25] Sun Zhaoxu , Han Min et al, " Multi-criteria Decision Making Based on PROMETHEE Method," 2010 International Conference on Computing, Control and Industrial Engineering.
- [26] Yongsheng Fang, Jun li , "A Review of Tournament Selection in Genetic Programming," Cai et al. (Eds.): ISICA 2010, LNCS 6382, pp. 181-192, 2010. © Springer-Verlag Berlin Heidelberg 2010.
- [27] Artem Sokolov, Darrell Whitley, "Unbiased tournament selection," Genetic and Evolutionary Computation Conference, GECCO 2005, Proceedings, Washington DC, USA, June 25-29, 2005.
- [28] R. Tinos, A.C.P.L.F. de Carvalho, "A genetic algorithm with gene dependent mutation probability for non-stationary optimization problems," IEEE, Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No. 04TH8753).
- [29] Jürgen Hesser, Reinhard Männer, "Towards an optimal mutation probability for genetic algorithms," Genetic Algorithms Genetic Algorithm Theory, International Conference on Parallel Problem Solving from Nature, Springer, PPSN 1990: Parallel Problem Solving from Nature pp 23-32.
- [30] Baswaraju Swathi, Harshvardhan Tiwari, "Test Case Generation Process using Soft Computing Techniques," International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278- 3075, Volume-9 Issue-1, November 2019.
- [31] Boyuan Chen et al, "An Automated Approach to Estimating Code Coverage Measures via Execution Logs," 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE).
- [32] Matina C. Donaldson-Matasci, Carl T. Bergstrom, and Michael Lachmann, "The fitness value of information," *Oikos*, PMC 2015 Apr 3.

Iterative Decoding of Chase Pyndiah Decoder Utilizing Multiple Relays Network

Saif E. A. Alnawayseh

Department of Electric Engineering
Faculty of Engineering, Mutah University
Karak, Jordan

Abstract—In this paper, a distributed Encoding and decoding of Turbo Product Code (TPC) over single and multiple relays network are proposed. The information message matrix is encoded at source by Bose Chaudhuri Hochquenghem (BCH) as component code and transmitted to destination and relays in the midway between source and destination. The coded source message is decoded by simple chase II decoder at relay and encoded again horizontally and vertically by BCH component code to construct TPC. Two scenarios were investigated. First scenario, utilizing one relay in cooperative network, where the vertical parity part of TPC is transmitted to destination to be the input of row decoder for original chase pyndiah decoder, while the received encoded horizontally matrix from source is the input of the column decoder. In the second scenario multiple relays are utilized and multiple copies of vertical parity part of TPC are sent to destination to be decoded by the proposed modified iterative chase pyndiah decoder with multiple integrated stages for each iteration. Simulation results for the first scenario over Additive White Gaussian (AWGN) and Rayleigh fading channels and using original chase Pyndiah decoder at destination shows 2dB gain improvement at BER= 10^{-5} and 4dB gain improvement at BER= 10^{-4} respectively over BER performance of un cooperative system. While results for distributed TPC decoding for the second scenario and using the proposed modified iterative chase Pyndiah decoder at destination shows 2.7 dB and 3 dB gain at BER = 10^{-4} for AWGN and Rayleigh fading channels respectively over the first scenario.

Keywords—Turbo product code (TPC); modified iterative chase Pyndiah decoding algorithm; relay; source; Bit Error Rate (BER); vertical parities; horizontal parities

I. INTRODUCTION

Distributed decoding of turbo codes has been investigated widely over cooperative relay network system [1], [2], [3], [4], and [5]. One of the turbo codes types is the Turbo Product Code TPC. TPC is considered to be an efficient coding algorithm compared to convolutional code turbo coed [6] which has been adopted by different communication standards such as IEEE802.20 and IEEE802.16 [7]. Many characteristics of TPC lead it to be one of the main coding schemes such as using simple Soft Input Soft Output (SISO) decoding algorithms that can reach near Shannon capacity limit by acceptable level of complexity of decoding especially at high code rates [8] such as iterative Pyndiah decoder [9] and many other decoding algorithms as in [10]. Moreover, TPC is easily constructed by components code which gives flexibility and high degree of parallelized structure and simple encoding. This simplicity and flexibility in encoding and decoding process can

be employed widely in coded cooperative relay communication system. Such system has gained a lot of interest recently as the diversity it provides against fading for wireless networks [11] and [12]. Furthermore, relay cooperative system is a good diversity alternative for multiple antennas where the size is a matter especially for wireless sensor networks [11]. The main idea of this system is based on the fact that inter user channel between source and relay or relay and destination are assumed to be more reliable than direct link between source and destination.

Many cooperative techniques are implemented according to this idea such as amplify and forward (AF), Decode and Forward (DF), and coded cooperation [13], [14], [15], and [16]. In this field coded cooperative using TPC shows a high performance with low complexity compared with conventional distributed codes [17] for its powerful BER performance and high rate code. Many works and ideas have been investigated to utilize TPC in single or multiple relays network. In [17] and [18] distributed TPC was investigated by sending BCH code to relays and destination. Where the received sequences at relays are decoded then encoded horizontally and interleaved circularly and encoded vertically to produce multiple copies of vertical parities that are send to destination to be decoded by modified SISO Turbo decoder. For maximum transfer information from relay a soft parity generating method that produces soft incremental redundancy was proposed in [19] using chase II soft decoding algorithm and distance based decoding. A further improvement for distributed TPC was presented in [20] using power allocation technique rather than assigning equally power to source and relay as done in [17] and [19]. Such method reduced the effect of relay decoding error compared to distribute TPC with fixed power assignment.

In this paper a distributed TPC is proposed by applying the idea of distributed encoding and decoding over single and multiple relay networks. Where the source message is distributed encoded between sources and relays to produce TPC, and distributed decoding over multiple channels between relays and destination utilizing modified iterative chase pyndiah decoder for TPC at destination. For efficient data transmission between relays and destination the source coded message matrix is decoded by chase II algorithm at relay and re-encoded gain horizontally and vertically using BCH component codes. The vertically parity encoded matrix is sent to destination to be the input of row decoder while the encoded message matrix from source is applied to the column decoder of iterative chase pyndiah decoder. For the case of multiple

relays, the modified iterative chase pyndiah will contain integrated multiple stages for each half iteration that capable to receive multiple versions of vertical parity matrixes from multiple relays. The main difference in the modified chase pyndiah decoder in this work compared to previous work in literature such as in [17] is that the extrinsic of the column decoder w_{ic}' in the first sub stage will be the input extrinsic for the next row decoder in the second sub stage. Also the Extrinsic of row decoder of second sub stage w_{ih}' will be the input for columns decoder. As a result, more efficient utilization for iterations in chase pyndiah iterative decoding as will be explained in details in next sections.

II. SYSTEM MODEL

We consider BCH systematic linear block codes C_1, C_2 with parameters (N_1, K_1, d_1) and (N_2, K_2, d_2) , where $N_i, K_i, and d_i$ are codeword length, input information block length, and minimum hamming distance for the code C_i respectively. The complete TPC matrix can be obtained by placing the $(K_1 \times K_2)$ data in Array of K_2 rows and K_1 columns and encoded the K_2 rows using C_1 and finally encode the N_1 columns using C_2 to produce N_2 , so the result will be TPC with $(K_1 \times K_2, N_1 \times N_2, d_1 \times d_2)$ as shown in Fig. 1. For this paper we assume the two component codes $C_1, and C_2$ are identical and have the same parameters $N_1 = N_2, K_1 = K_2, and d_1 = d_2$. In below paragraphs a summary of the steps adopted in this paper:

A. Distributed Coded System

To establish the distributed encoding for the TCP, we implement two scenarios the first with single relay and the second with double relays. In which relay nodes are located in the midway between source and destination. The source (S) broadcast the $(K \times N)$ codewords matrix encoded by C to the destination (D) and the relays (R) that are located in the midway between source and destination as shown in Fig. 2. Separations between Rs and D are as in [21] and [22], where the distance between S and D is normalized. The transmitted codeword $C = C_1, C_2, \dots, C_N$ is modulated by Binary Phase Shift Keying BPSK to get $X = X_1, X_2, \dots, X_N$. So the received signal at R and D in the first time slot are:

$$r_{sdi} = \alpha_{sdi} h_{sdi} X_i + W_{sdi} \quad (1)$$

$$r_{sri} = \alpha_{sri} h_{sri} X_i + W_{sri} \quad (2)$$

Where $X_i, 1 < i < N$ is the modulated transmitted horizontally codeword from source, and h_{sdi}, h_{sri} and W_{sdi}, W_{sri} are the fading coefficients and AWGN of S-D and S-R links respectively. And $\alpha_{sdi}, \alpha_{sri}$ are the path loss attenuation for the two links affected by distance between nodes $\alpha_{sdi} = 1$, and $\alpha_{sri} = (1 - \lambda)^{-\mu}$ where μ is the path loss exponent and λ is the distance between relay and destination. The received sequence at R is decoded by Chase II SISO decoder as in [23] and the detected information words K_1 are re-encoded horizontally by component code C and encoded vertically to produce vertical parity part V_p as shown in Figure 1. In the second time slot the coded vertically part V_p is transmitted from R to D and received as:

$$r_{rdi} = \alpha_{rdi} h_{rdi} X'_i + W_{rdi} \quad (3)$$

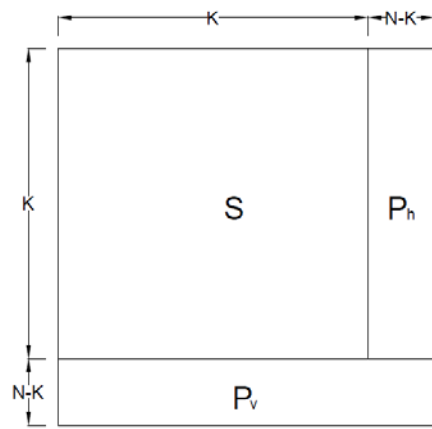


Fig. 1. The structure of TPC.

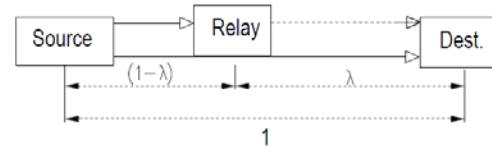


Fig. 2. Cooperative Model with Three Terminals.

Where $X'_i, 1 < i < N$ is the modulated BPSK symbols of vertical parity column V_p , and h_{rdi}, W_{rdi} are the fading coefficients and AWGN of R-D link. It should be noticed that the transmitted power from S and R are equal and path loss exponent μ is 2. So we obtain the values of:

$$SNR_{sr} = \frac{SNR_{sd}}{(1-\lambda)^2}, \quad SNR_{rd} = \frac{SNR_{sd}}{(\lambda)^2} \quad (4)$$

From (4) the variance of W_{sr} and W_{rd} in (1,2 and 3) can be expressed as: $\sigma_{sr}^2 = (1 - \lambda)^2 \sigma_{sd}^2, \sigma_{rd}^2 = (\lambda)^2 \sigma_{sd}^2$ Where the variance of $W_{sd} = \sigma_{sd}^2, W_{rd} = \sigma_{rd}^2$, and $W_{sr} = \sigma_{sr}^2$.

B. Distributed TPC Decoding using Single Relay

The two parts of TPC matrix $\{S, P_h\}$ and $\{P_v\}$ are received in D at the second time slot via direct link and relay channel with SNRs as in (1) and (3). At D the received $\{S, P_h\}$ and $\{P_v\}$ are arranged to construct TPC as in Fig. 1. In this paper iterative chase pyndiah decoder as shown in Fig. 3 is used as SISO decoder to decode the received $\{P_v\}$ part from R by column decoder and received rows $\{S, P_h\}$ through direct link by row decoder. Soft decoding for $\{P_v\}$ in columns decoder is performed using chase algorithm as following:

- 1- The hard decisions vector $R_d = (r_{d1}, r_{d2}, \dots, r_{dN})$ is generated from the soft received sequence $r_{rd} = r_{rd1}, r_{rd2}, \dots, r_{rdN}$ according to:

$$r_{rdi} = \begin{cases} 1, & \text{if } r_{rdi} \geq 0 \\ 0, & \text{other wise} \end{cases}$$

Where $i \in \{1, 2, \dots, N\}$

- 2- The reliability of r_{rdi} component $|r_{rdi}|$ is ordered and positions of P of Least Reliable Bits LRB are determined, then 2^P test patterns T_l are obtained by placing combinations of ones and zeros at P LRB and zeros in the remaining positions.

- 3- The test sequence Z_l is obtained by modul-2 addition between T_l and R_d as:

$$Z_l = R_d \oplus T_l$$

Where $l \in \{1.2 \dots 2^P\}$

- 4- Syndrome is calculated for each Z_l by $S_l = Z_l \cdot H^T$, where $l \in \{1.2 \dots 2^P\}$ and H^T is the transposed parity check matrix of the component code C. subsequently error correct based syndrome is conducted to generate valid codeword $C^l = (C_1^l, C_2^l, \dots, C_N^l)$

- 5- Squared Euclidian Distance (SED) is calculated between R_d and C_l according to :

$$|r_d - C^l| = \sum_{i=1}^N [r_{di} - (2C_i^l - 1)]^2 \quad (5)$$

- 6- The best candidate word D is chosen by calculating the minimum SED among the 2^P candidate codewords by:

$$D = \operatorname{argmin}(C^l, |R_d - C^l|) \quad (6)$$

- 7- In the present of competing codeword D related to the i_{th} bit position the extrinsic information is calculated for the column decoder as:

$$w_{ic} = \frac{\|r_d - C^{l(K)}\|^2 - \|(r_d - D)\|^2}{4} (2d_i - 1) - r_i \quad (7)$$

Where d_i is the i_{th} element decision codeword D and $C^{l(i)}$ is the competing codeword with minimum SED among the candidate codeword carrying value different from the i_{th} of $C^{l(K)}$. if this candidate codeword not exist w_i is calculated as $w_i = \beta(2d_i - 1)$, where β is the reliability factor and presented as in [9].

- 8- Once the extrinsic information is determined the input of next decoding stage (rows decoder) is inserted as follow:

$$r'_i = r_{sd i} + \alpha w_i \quad (8)$$

Where α is the wight factor as in [1in parallel] to combat high standard deviation in extrinsic w_i and high BER during first iteration. And $r_{sd i}$ is the received sequence from direct link which contain horizontally coded words $\{S, P_h\}$.

All operations above are repeated on all bits of TPC codewords from step 1 to 8 to achieve the first half iteration as shown in Fig. 3 where the integer m is the half iteration.

C. Distributed TPC Decoding using Multiple Relays

In this paper only two relays are considered for demonstration cooperative system with multiple relays. The S broadcast the block $\{S, P_h\}$ to Rs and D. The two relays act independently, the two received sequence at Rs are decoded by chase Π SISO decoder as in [23] and the detected information blocks are encoded again horizontally and vertically by component coded C to construct at TPC as shown in Fig. 1. In the second time slot the encoded vertical block $\{P_v\}$ are transmitted from both relays to destination as:

$$r_{1rd i} = \alpha_{rd1 i} h_{rd1 i} X'_{rd1 i} + W_{rd1 i} \quad (9)$$

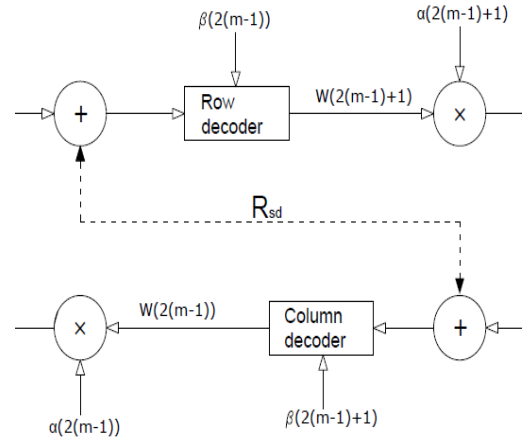


Fig. 3. Original Iterative Chase Pyndiah Decoder.

$$r_{2rd i} = \alpha_{rd2 i} h_{rd2 i} X'_{rd2 i} + W_{rd2 i} \quad (10)$$

Where $X'_{rd1 i}$ and $X'_{rd2 i}$, $1 \leq i \leq N$ are the modulated vertically coded symbols from the two relays. $h_{rd1 i}$ and $h_{rd2 i}$ are the fading coefficients for the first and second relay links to destination respectively. The vertically parity coded parts are received from R and horizontally codewords from S through direct link are $\{P_{v1}, P_{v2}, P_h, S\}$. For each generated vertical parities $\{P_{v1}, P_{v2}\}$ there are two decoding stages at D: first through rows and then through columns decoder. So in this paper the proposed modified chase pyndiah decoder that contain multiple integrated sub stages for each iteration as shown in Fig. 4 is used. The first sub stage receives r_{1rd} sequence the contains vertical parties columns $\{P_{v1}\}$ from the first relay are inserted to the first row decoder in sub stage 1 which perform the decoding process as explained in previous subsection for single relay case with steps from 1to 8. The main difference in the modified chase pyndiah decoder in this work is that the extrinsic of the column decoder w_{ic}' in the first sub stage will be the input extrinsic for the next row decoder in the second sub stage and can be written as

$$w_{ic}' = \frac{\|r_{sd i} - C^{l(K)}\|^2 - \|(r_{sd i} - D)\|^2}{4} (2d_i - 1) - r_{sd i} \quad (11)$$

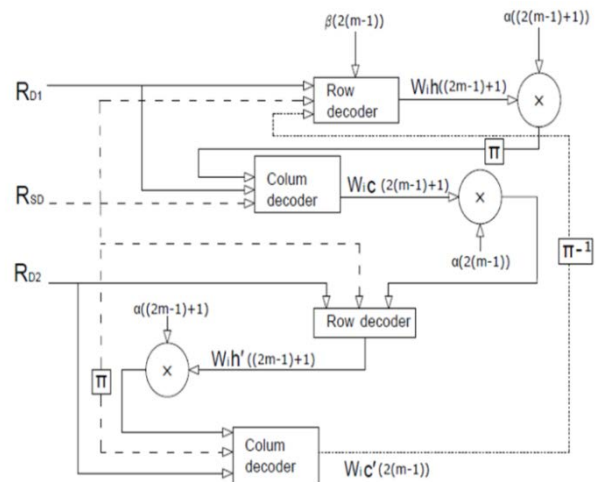


Fig. 4. Modified Iterative Chase Pyndiah Decoder.

And the input sequence for the row decoder in the second sub stage will be:

$$r1' = r_{2rdi} + \alpha w'_{ic} \quad (12)$$

The rows decoder in the second sub stage will implement the decoding procedures for single relay case following steps from 1 to 8. Extrinsic of row decoder of second sub stage w'_{ih} will be the input for columns decoder exactly as in first sub stage.

$$r2' = r_{sdi} + \alpha w'_{ih} \quad (13)$$

III. RESULTS AND DISCUSSION

In this section the simulation results obtained for distributed TPC utilizing chase pyndiah decoder for single relay and multiple relays are shown. In this paper a squared (15,11,3) hamming code based TPC with code length 225 is used. For simulation positions of $P = 10$ of Least Reliable Positions LRB are determined. The simulation is carried over AWGN so the block fading coefficients in (1), (2), and (3) are normalized to 1. Also the relays are assumed to be centered at the distance between source and destination. The error correcting performances will be represented in terms of BER and Signal to Noise Ratio ($SNR=E_b/N_0$). The BER of TPC (15.11.3)² code over a single link AWGN (un-cooperative) channel using chase pyndiah decoder at destination is shown in Fig. 5. It can be observed that the fifth iteration reaches $BER=10^{-5}$ at 6 dB while reaches $BER=10^{-3}$ at 4 dB with 3 dB gain over uncoded system. In Fig. 6, the BER the performance results are shown for the same TPC but over Rayleigh fading channel. The fifth iteration reaches $BER=10^{-3}$ at 15 dB with more than 6 dB gain compared to uncoded BPSK system.

A. For Single Relay

The simulation results for TPC coded system utilizing single relay in cooperative system for free space propagation and the path loss exponent $\mu = 2$ over AWGN and Rayleigh fading channels utilizing distributed coding and decoding with iterative chase pyndiah decoder at destination are shown respectively in Fig. 7 and Fig. 8.

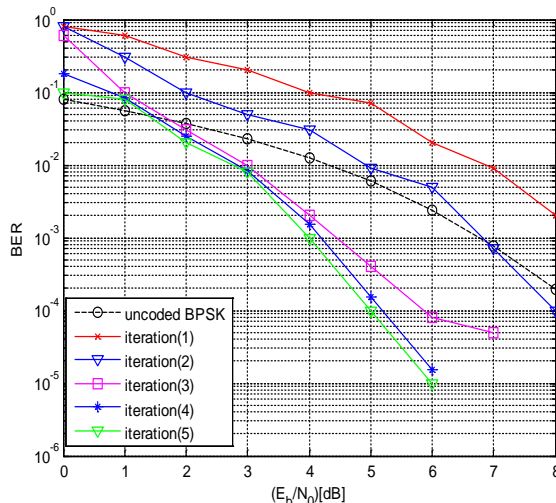


Fig. 5. The BER of the TPC over an AWGN Channel without Relays.

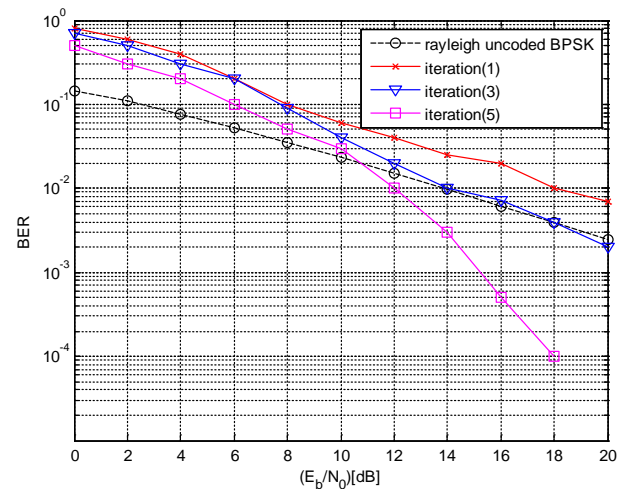


Fig. 6. The BER of the TPC over Rayleigh Fading Channel without Relays.

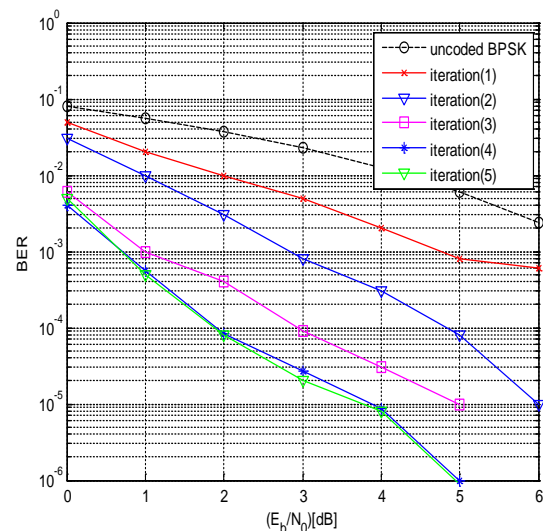


Fig. 7. BER Performance of Distributed TPC utilizing Single Relay over AWGN.

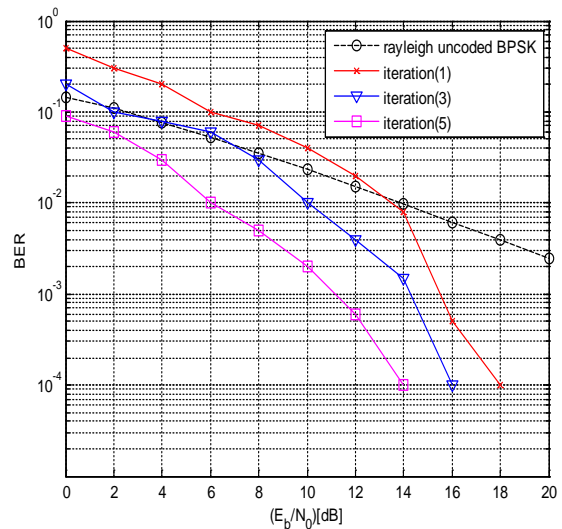


Fig. 8. BER Performance of Distributed TPC utilizing Single Relay over Rayleigh Fading Channel.

The BER performance for distributed TPC utilizing single relay and iterative chase pyndiah over AWGN is shown in Fig. 7. The proposed distributed TPC decoder at destination with vertical columns parity $\{P_v\}$ from the relay is the input of rows decoder and the horizontally codewords from source $\{S.P_h\}$ are the input of columns decoder. The fifth Iteration reaches 10^{-5} BER at SNR=4 dB with coding gain about 2 dB compared uncooperative TPC over AWGN in Fig. 5. In Fig. 8 the BER performance for distributed TPC utilizing single relay and iterative chase pyndiah over Rayleigh fading channel is shown. With same distributed decoding for TPC as in the system in Fig. 8 the fifth iteration reaches 10^{-4} BER at 14 dB SNR with coding gain about 4 dB compared to uncooperative case TPC in Fig. 6 for the same system parameters.

B. For Multiple Relays

In this paper only two relays are considered to construct the cooperative system. Where the proposed distributed TPC decoder at destination receives the vertical columns parity $\{P_{v1}\}$ from the first relay as the input of row decoder and the horizontally codewords from source $\{S.P_h\}$ are the input of columns decoder in the first sub stage decoding. While receives the vertical columns parity $\{P_{v2}\}$ from the second relay as the input of rows decoder and the horizontally codewords from source $\{S.P_h\}$ are the input of columns decoder in the second sub stage decoding. And the extrinsic of the columns decoder w_{i_c}' in the first sub stage will be the input extrinsic for the next row decoder in the second sub stage of the modified iterative chase pyndiah decoder.

We compare the performance of distributed TPC using modified iterative chase pyndiah decoder utilizing two relays with distributed original iterative chase pyndiah decoder utilizing one relay over AWGN and Rayleigh fading channel. In Fig. 9 the fifth iteration over AWGN reaches BER = 10^{-5} at SNR =2.7 dB approximately while the distributed TPC iterative chase pyndiah with single relay at Fig. 7 reaches the same BER at SNR =4dB while the uncooperative reaches the same BER at SNR=6dB. So it can be observed that modified chase pyndiah decoder with 2 relays achieves 1.3 dB gain over single relay and 3.3dB gain over un-cooperative system.

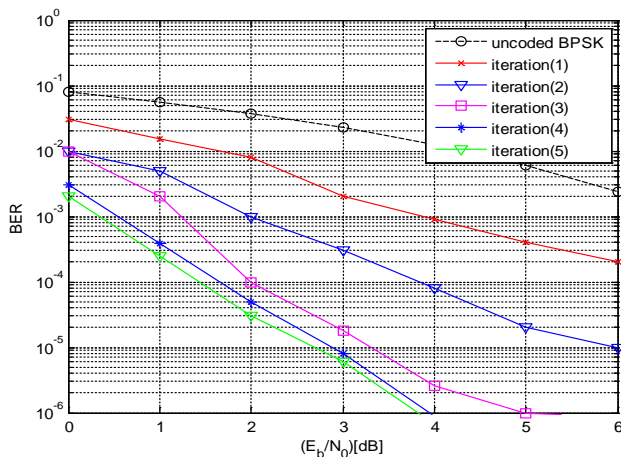


Fig. 9. BER Performance of Distributed TPC utilizing Multiple Relays over AWGN.

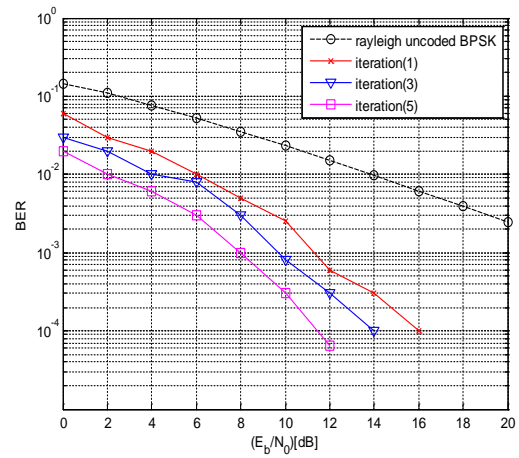


Fig. 10. BER Performance of Distributed TPC utilizing Multiple Relays over Rayleigh Fading Channel.

For Rayleigh Fading channel in Fig. 10 with same system parameters of Fig. 9 the distributed decoding with iterative modified chase pyndiah decoder shows coding gain of 3 dB for the fifth iteration compared single relay scenario at BER= 10^{-4} in Fig. 8 and here it can be noticed the efficiency of increasing number of iterations with proposed modified chase pyndiah iterative decoder.

IV. CONCLUSION

The results presented in this paper illustrate the distributed coding and decoding of TPC using iterative chase pyndiah decoder over single and multiple relay networks. Two scenarios were investigated for distributed coding and decoding , the first utilizing one relay in cooperative network where the original chase pyndiah decoder at designation receives $\{P_v\}$ vertical parity codewords from relay as the input of rows decoder and the horizontally codewords $\{S.V_h\}$ from source as input to column decoder. In the second scenario multiple relays are utilized and multiple copies of vertical parity parts of TPC are sending to destination to be decoded by modified iterative chase pyndiah decoder with multiple integrated stages for each iteration. Simulation results for proposed system with single relay and multiple relays modified iterative chase Pyndiah decoder show gain improvement respectively over BER performance of uncooperative system.

REFERENCES

- [1] Zhao and M. C. Valenti, "Distributed turbo coded diversity for relay channel," Electronics Letters, vol. 39, no. 10, pp. 786-787, 2003.
- [2] M. C. Valenti and Bin Zhao, "Distributed turbo codes: towards the capacity of the relay channel," in IEEE 58th Vehicular Technology Conference, VTC, Orlando, 2003.
- [3] K. Anwar and T. Matsumoto, "Accumulator-Assisted Distributed Turbo Codes for Relay Systems Exploiting Source-Relay Correlation," IEEE Communications Letters, vol. 16, no. 7, pp. 1114-1117, July 2012.
- [4] Mughal, S.; Yang, F.-F.; Ejaz, S, " Asymmetric turbo code for coded cooperative wireless communication based on matched interleaver with channel estimation and multi-receive antennas at the destination,," Radio engineering , vol. 26, pp. 878-889, 2017.
- [5] C.Zhao, F. Yang, R. Umar and S. Mughal, "Two-Source Asymmetric Turbo-Coded Cooperative Spatial Modulation Scheme with Code,," electronics, vol. 9, no. 1, pp. 1-20, 2020.

- [6] C. Argon and S. W. McLaughlin, "An efficient Chase decoder for turbo product codes," *IEEE Transactions on Communications*, vol. 52, no. 6, pp. 896-898, June 2004.
- [7] C. Xu, Y. Liang and W. S. Leon, "A Low Complexity Decoding Algorithm for Turbo Product Codes," in *IEEE Radio and Wireless Symposium*, Long Beach, CA, 2007.
- [8] J. Cho and W. Sung, "Reduced complexity Chase-Pyndiah decoding algorithm for turbo product codes," in *IEEE Workshop on Signal Processing Systems (SiPS)*, Beirut, 2011.
- [9] R. M. Pyndiah, "Near-optimum decoding of product codes: block turbo codes," *IEEE Transactions on Communications*, vol. 46, no. 8, pp. 1003-1010, Aug 1998.
- [10] S. Yoon, B. Ahn and J. Heo, "An advanced low-complexity decoding algorithm for turbo product codes based on the syndrome," *EURASIP Journal on Wireless Communications and Networking*, vol. 126, no. 16, pp. 1-31, 2020.
- [11] A. Sendonaris, E. Erkip and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927-1938, Nov.2003.
- [12] H. Liang ,A. Liu, J. Dai, J. Liu, and C. Gong, "Design and analysis of polar coded cooperation with incremental redundancy for IoT in fading channels," *IET Communications*, vol. 15, no. 4, pp. 595-602, 2021.
- [13] R. Hu and J. Li, "Practical Compress-Forward in User Cooperation: Wyner-Ziv Cooperation," in *IEEE International Symposium on Information Theory*, Seattle, WA, 2006.
- [14] H. H. Sneessens and L. Vandendorpe, "Soft Decode and Forward Improves Cooperative Communications," in *2005 6th IEE International Conference on 3G and Beyond*, Washington, DC, 2005.
- [15] Y. Li, B. Vucetic, T. F. Wong and M. Dohler, "Distributed Turbo Coding with Soft Information Relaying in Multihop Relay Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 11, pp. 2040-2050, Nov. 2006.
- [16] Y. Hairej, A. Darmawan and H. Morikawa, "Cooperative Diversity using Soft Decision and Distributed Decoding," in *2007 16th IST Mobile and Wireless Communications Summit*, Budapes, 2007.
- [17] E. A. Obiedat, G. Chen and L. Cao, "Distributed Turbo Product Codes over Multiple Relays," in *2010 7th IEEE Consumer Communications and Networking Conference*, Las Vegas, 2010.
- [18] E. A. Obiedat, L.Cao, "Cyclic Interleaving for Multiple Vertical Parities in Distributed Turbo Product Codes," *Recent Patents on Signal Processing*, vol. 4, no. 1, pp. 12-17, 2014.
- [19] E. A. Obiedat and L. Cao, "Soft Information Relaying for Distributed Turbo Product Codes (SIR-DTPC)," *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 363-366, April. 2010.
- [20] E. A. Obiedat and L. Cao, "Power Allocation for Distributed Turbo Product Codes (DTPC)," in *IEEE Global Telecommunications Conference GLOBECOM 2010*, Miami, FL, 2010.
- [21] S. E. A. Alnawayseh and P. Loskot, "Cooperative Versus Receiver Coded Diversity with Low-Complexity Encoding and Decoding," in *2010 IEEE 71st Vehicular Technology Conference*, Taipei, 2010.
- [22] S.E. A. Alnawayseh, "Performance Comparison of Coded OFDM System with Cooperative Diversity and Multi- Antenna Receiver Diversity using QAM modulation," *JJEE Jordan Journal of Electric Engineering*, vol. 1, no. 2, 2015.
- [23] D. Chase, "Class of algorithms for decoding block codes with channel measurement information," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 170-182, 1972.

An Efficient Privacy Preserving Approach for e-Health

Supriya Menon M¹

Research Scholar

Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddesvaram, AP, India

Dr. Rajarajeswari Pothuraju²

Professor

Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddesvaram, AP, India

Abstract—Immense Procreation of large amounts of data in medical field and health care domain, benefitting society is at risk with sensitive attributes being disclosed. Access to Medical Information made feasible over internet with an intension of serving the people related to medical community is triggering a challenge for researchers in norms of Privacy and security. The medical data at cloud is vulnerable to unpredictable threats with evolving technology, and the threat landscape sounds resilient with sensitive attributes. In this contemporary stretch, Organizations fail to hold the reputation and are unable to preserve public confidence. The austerity of sophisticated security attacks compromise the privacy of patient data and security of healthcare units. The fruitful approaches by several researches and practitioners provided an up heal resolutions, but the demand for an optimal solution remains unanswered. In this paper we present a solution for addressing the security issues in health care management. We propose a hybrid framework using enhanced Attribute Based Encryption (ABE) with Anonymity approach based on access primitives of sensitive attributes. The proposed mechanism is evaluated in terms of performance, encryption time, decryption time and Memory utilization using Jsim simulator which envisage drastic performance expedition in the presented model.

Keywords—e-Health; Attribute Based Encryption (ABE); secure hash algorithm (SHA-1); anonymity; privacy; sensitive parameters

I. INTRODUCTION

Medical domain owes to be attractive region for researchers, challenging them in various aspects like disease prediction, Drug prediction, Drug Repositioning and many more. The recent research focused on disease and treatment prediction using medical repositories accessed and published over distributed environments [1]. This Medical records accessed by several authorities put forth's queries questioning the security and privacy of health care data. The cause entailing such need divulged from the fact that health care data is outsourced for various reasons at the risk of compromising privacy requirements like confidentiality [2, 3], Integrity, keyword privacy, authentication audit ability and further. Additionally evolving technologies of information and communication attracts medical domain for integrating health data with technology [4] from domain's like Hospitals, health insurance firms and research laboratories leading to e-Health.

E- Health, an attractive domain in recent times that overlaps public health and medical informatics with corporate sectors over internet aims at improving data analysis of health care data locally and worldwide. The cloud successfully offers few advantages over network primarily in enhancing patient care by supporting interaction [5] with healthcare authorities and availability of patient data for analysis and diagnosis [6]. e-Health also offers support for medical research in disease treatment prediction with extended monitoring of epidemics. Further, it helps in cost reduction for engaging expensive hardware, software and data storage at premises.

Eventually e-Health showcases few pullbacks like complexity in interoperability i.e. lacking standard for synchronization, security and privacy issues [20] in shared and public environments [8], regulation controversies related to social and valid frameworks and reliability considerations. They even need to work hands with sensor networks [9] involving data collection. The Healthcare providers in practical are surrounded with several risks [10] from digital technology on cloud despite they encircled advantages.

Among aforementioned issues security and privacy challenges demand utmost attention for realization of its effective utilization. Data confidentiality, authentication and Integrity are at risk in distributed environment. The goal of medical data shared and stored over internet is to provide consistency and high level of security. Despite numerous cryptographic and non-cryptographic methodologies available for enduring security and privacy of e-health data [11, 12], few unturned grains are hindering hurdles constraining performance. Our approach proposes a way forward for contending the security and privacy gaps in e-health [13].

The proposed architectural framework attains the goal of security using a Hybrid ABE as well as provides selective access to records based on user predefined access policies like authorized uses, restricted users and un authorized users. The Hybrid ABE provides efficient performance using secure hash algorithm (SHA-1) and Anonymity approaches. The hybrid approach promises high degree of availability, reliability, and efficiency in protecting patient sensitive information [14] upon implementation. Explicitly the architecture gives room for desired authentication to medical archives with extended control for medical stakeholders in general and emergency scenarios on demand.

II. RELATED WORK

Puneeth Saran et al 2020 contributed to a qualitative research regarding role of cloud computing in medical field and proposed a method to increase the security of medical records in cloud. Gaozhiqiang et al 2015, proposed a cloud based remote health care consisting of portable medical devices, intelligent terminals, cloud platforms where user can access their health data via internet. Inderpreet Singh et al 2019, presented a model for grouping adaptable e-health care services depending on distributed computing environment which showcases high correctness rate for secure information access. KnutHaufe et al 2014 presented a framework for security of health care records stored in cloud and identified ISMS process that needs to be focused for future research. R.Anitha & Saswathi Mukherjee 2014 proposed a novel framework -generating cipher key from attributes of metadata created by DCMI standard using patients medical record. Luliana Chiuchisan et al 2017 made a detailed survey of security measures involved in health care management and proposed health care system that monitors rehabilitation of patients with Parkinson's disease. Alexandru Soceanu et al 2015 presented encryption scheme and attribute based framework with encryption process relying on ARCANA tool for secure hierarchical access. Yaza-Al-Issa et al. 2019 reviewed with regard to cloud computing services in health care management and privacy concerns for health care providers and reiterated that only few concerns of security are addressed. Shekha Chentharra et al. 2019 contributed to intensive survey about HER (Electronic Health Record) security and privacy, EHR cryptographic and non cryptographic approaches in IEEE, Science direct, Google scholar, Pub Med and ACM library. Nureni Ayofe Azeez & Charles Van Der Vyer, 2018 reviewed 110 original articles, figured out various models adopted with their standard definitions on e-health and proposed secured architecture for e-health to provide privacy between health care providers and patients. Isma massod, 2018, proposed six step generic framework for patient physiological parameters, privacy and security in sensor supported cloud infrastructure with performance evolution in research. Ronald Glasberg et al., 2014, analyzed risks and crisis for health care providers in holistic way, taking organizational and human aspects into account. Shyh -Wei Chen et. al., 2016, proposed architecture of patient centered personnel health record to manage patient health information and health reports with cloud based secure transmission. Alexmu-Hsing Kuo et. al., discussed in detail about health care and considers four aspects to analyze the challenge of cloud computing model. Ramzi. A et.al. presented recovery algorithm using concept of matrix in health care management and evaluated the performance against various techniques. Panjunsun, 2019, proposed privacy protection framework by reviewing challenges and solutions of data security in detail. Uma narayanan et al. 2020, proposed novel system architecture called security authentication and data sharing in cloud (SADS –cloud) including SHA-3 hashing algorithm for registered data owners. Ijaz Ahmad Awan, 2020 proposed framework deploys AES with 16, 32, 64, and 128 plaintext bytes enhancing security and minimizing resource utilization in computational clouds. Arafat Al-Dhaqm et al., 2020, gave a detailed review on DBFI –

Database Forensic Investigation and proposed harmonized DBFI process using systematic approach with higher certainty. SupriyaMenon M and Rajarajeswari P, 2018, reviewed privacy issues of personalized and context aware privacy and proposed a model for context aware privacy. Jitendra Kumar and Ashutosh Kumar Singh, 2017, came up with a workload prediction model using Long short term memory (LSTM) and tested over three web log datasets proving enhanced accuracy by proposed approach. Supriyamenon M and Rajeswari P, 2020, addressed the complications related to drug repositioning and came up with a hybrid ACO approach enhancing Drug consumption similarities for better repositioning addressing the need for secure patient data. Ma, H., Zhang, R., & Yuan, W, 2016, contributed a model for ABE based Anonymity for Identity revelation.

III. SECURITY PRELIMINARIES

A. ABE

ABE is an encryption scheme, where the generated cipher text is an outcome relying on user private key and attributes of user data. This public key encryption technique renders plaintext at requested site with decryption supported upon attribute matches of user key and cipher text attributes [15] from attributes of metadata. Although initially introduced in its basic form, exploring amendments of attribute based encryption [16] with multiple authorities involving in user private key generation are also available. ABE has its wide spread usage in several areas like vector driven search engine interfaces, log encryption avoiding log encryption with all recipient keys.

There are two forms of ABE one for key policy KP-ABE and other for cipher policy CP-ABE. The KP-ABE generates user private depending on access tree related to user privileges and encrypting over a set of attributes using algorithms like AES [17]. However cipher text based ABE encrypts user data and attribute with secret keys generated from access trees.

ABE rising to be a well preferred mechanism is surrounded with overwhelming challenges like in efficient attribute revocation mechanism, improper key co-ordination, key escrow deficiency and issues related to key revocation mainly for healthcare systems [18]. Few extended problems in the path of ABE is its centralized concept. The need for a centralized body or authority participating in private key generation, makes ABE encounter the flaws due to lack of decentralization. These risks bring down the performance of ABE. One more factor of concern affecting the ABE is speed, which downtrends compared to others due to delay of policy tree construction and computational delay at decryption site also adds upon the issue.

B. SHA-1

The secure hash algorithms enable the determination of Message Integrity that facilitate creation and validation of digital signatures. Digital signatures provide secure security service of Authentication [19] hereby avoiding Denial attacks and repudiations both at source and destination. SHA-1 belongs to the family of secure hash algorithms that generate a hash value known as message digest to facilitate security [21]. It promises its wide spread excellence in several security

protocols, mail protocols, TTL, SSL, IPsec and many more. The basic version of the algorithm produces a 160 bit message digest which well prevails against Brute force attack. This variant is considered to be the fastest one but more prone to collision problem, those were overruled in the successor variants. Few well known variants of secure hash algorithms are SHA -2 and SHA-3. The former uses a set of 6 hash functions with digests of size 224, 256, 384, and 512 bits. Among the aforementioned digests SHA-256 and SHA 512 exhibit uniqueness in the sense of computing with 32 bit and 64 bits respectively. They project the variation in the basic shift and additive operations performed. SHA-2 being advanced faced a strong battle to take over its Predecessor. The later addressed SHA-3 by NIST provides compatibility with the former.

C. Anonymization

Anonymization is a process that aims at encapsulating identifying information in a way intending privacy protection. Hence the original data remains anonymous enabling data sharing and transmission among agencies reducing risk of unwanted disclosures [22]. Despite such secure transformations anonymous data never promise to anonymous over time. Several approaches and clever techniques exist that disclose data leading to be de-anonymized. To handle all such loop holes, several forms of Anonymity are available like k-Anonymity, l-Anonymity, t- closeness, p-sensitivity and many more variations. In k-anonymity, anonymization is a key feature using certain cryptographic hashing. K-Anonymity further has its extension to an (α , k)-anonymity model for privacy preserving data publishing, where α being a fraction and k an integer. The frequency of sensitive value is no more than α . It aims at data security and privacy with further extension to Human and Societal aspects of security and privacy.

IV. PROPOSED APPROACH

Huge amounts of data filling the health care repository is triggering several challenges in due response to providing services. These services claim that cloud computing techniques provide everything as a service i.e. storage as well as security as a service. The major issue of concern is medical confidentiality, portraying the healthy relationship of trust among patients and doctors. The medical data stored in cloud is at high risk of being vulnerable to attacks with irretrievable loss to users with their sensitive data dumped at entrusted servers. With an intension of addressing the above mentioned issues related to data privacy we propose an hybrid approach that resolves the complications in data transmission and provide security.

Phases in proposed approach are discussed below.

Phase 1:

This phase of the proposed system initiates with generation of metadata for the patient records. The attributes in patient records are analyzed and access control structure is defined considering different threshold parameters for various groups of users using ABE approach. Certified attributes defined in the access policy determines which block of plain text should be decrypted for the users with predefined threshold

credentials. Elicited from the defined access policies, users are assigned access permissions to the available records.

Phase 2:

The medical records blocks are encrypted considering four randomized algorithms in ABE as Setup, Key generation, Encryption and Decryption.

Setup: At initialization the system generates 2 groups GR1 and GR2 based on security parameters with p prime value, t threshold and b bilinear pairs.

The centralized authority generates master key M_K and public key P_K by randomly selecting $x, u_1, u_2, \dots, u_n \in Z_q$ where q is the prime number and Z_q multiplicative modulo.

Key generation: The authorized authority generates secret key S_k for users by using SHA1 with modified feistel structure where SHA1 converts attributes into matrices considering m rows and n columns, where m is the number of attributes and n is the size of SHA output.

The algorithm for key generation is presented below.

1. Initiates by reading sensitive attributes and inputting them to SHA-512.
2. Resultant Matrix $A_{m \times n}$ is further divided into $A_1, A_2, A_3,$ and A_4 .
3. Produce $L_{m \times n}$ and $R_{i \times m}$ by combining A_1, A_3 and A_2, A_4 respectively.
4. Left and Right values of Feistel network undergo following computations.
 - Bifurcating $R_{i \times m}$ to equal partition matrices R_{ia} and R_{ib} .
 - Apply transposition resulting in $R_{ia \times m}$ and $R_{ib \times m}$ and add them to $Q_{m \times n}$.
 - $RO_{m \times n}$ a resulting transpose of $Q_{m \times n}$.
 - Revised $L_{m \times n}$ is $RO_{m \times n}$ and $R_{i \times m}$ is previous value of $L_{m \times n}$ until n holds old value.
 - Output $K_{m \times n} = L_{m \times n} \parallel R_{i \times m}$.
5. Process terminates.

Encryption: Sender encrypts the message with key extracted from attributes.

$C_i = \text{Encryption} (P_k, PT, A)$ where P_k public key from attributes A for plaintext PT.

Decryption: The Receiver decrypts Cipher C_i using the Secret key S_k generated by SHA1.

ABE explicitly supports threshold operations on attributes to specify permitted access control structures to the users of different groups.

Phase 3:

Among different groups of users with certain attribute combinations, the limited access groups of users considered as restricted users are subjected to feasible Anonymity technique

with lower distortion. The k- anonymity technique preferred avoids identity disclosure.

Algorithm for (a, k)Anonymity :

Input: Raw table

Output: Hybrid Anonymity table.

1. Initialization stage

Generate user for input vector and for array of users considered 1,2,3,...n, compute dissimilarity matrix DMT by calculating distances.

2. Anonymization at Client side

1. Compute DMT.

2. Assign false to all points and select a point p with C_c as centroid of user c_i and mark it as true.

3. Consider false points as minimum distance from C, with social attributes S_A .

4. Add above considered point to P, and check the frequency with Anonymization parameter k.

If frequency of (S_A) < k

```
{
Consider and adjust centroid;
}
```

Else

Abort;

4. Repeat until all points in c are verified and return.

5. Group unassigned points to nearest user and ensure user satisfying (1,K1) anonymity.

3. Anonymization at Server side

1. Consider the nearest user pairs p1 and p2 in client side matrix.

2. Combine p1 and p2.

3. Size of $u=p1+p2$.

4. Compute representative vector $R^* u$ using tree access structure T.

5. T_y denotes sub-tree where y is root in tree T.

When root node is r in $T = T_r$ Attribute set = e that confirms to access policy of T_y

Then $T_y(e) = 1$

If y is a leaf node and attribute attr (y)

$T_y(e) = 1$

Else

Validate sub- nodes.

6. Repeat process until each user satisfies (a2, k2) anonymity.

Based on the Anonymity levels attained from proposed technique predefined threshold attributes in access tree structure are sent in Plain text and other blocks are anonymized.

V. PERFORMANCE EVALUATION AND ANALYSIS

Our Proposed Hybrid ABE Approach projects efficient performance with respect to time and Memory Utilization when compared with existing techniques like Common

Database Forensic Investigation Process (CDFIP), Real-time Operational Data Base (RODB) Extraction–Transformation–Loading (ETL), and Long Short Term Memory (LSTM) [7] to Recurrent Neural Network (RNN). The Simulation parameters considered for Implementation of the proposed approach are shown in the Table I.

To evaluate the performance of Hybrid ABE, Encryption time, Decryption time and Memory Utilization are considered.

Encryption time: This computes the throughput of the encryption scheme with respect to user instances and encryption time.

Decryption time: This computes the throughput of the decryption scheme with respect to user instances and decryption time.

Memory Utilization: This evaluation parameter projects the average utilization of system memory in bytes for different user instances.

Table II shows an improved performance of time for different user instances using Hybrid Approach against existing methods opted.

Fig. 1 depicts improved performance of the Proposed Approach contradicting existing Approaches mentioned.

TABLE I. SIMULATION PARAMETERS

PARAMETERS	VALUES
Simulator	Jsim
Simulator Time	120 s
Proposed Protocol	Hybrid Approach
Number of user instances	10 - 500

TABLE II. PERFORMANCE COMPARISON

No. of user instances	CDFIP	RODB & ETL	LSTM to RNN	Hybrid Approach
10	4.3	3.7	3.6	1.9
30	5.4	4.8	5.2	3.4
50	6.4	5.4	4.6	3.7
70	7.3	6.2	7.1	4.6
100	8.6	7.4	6.3	7.3

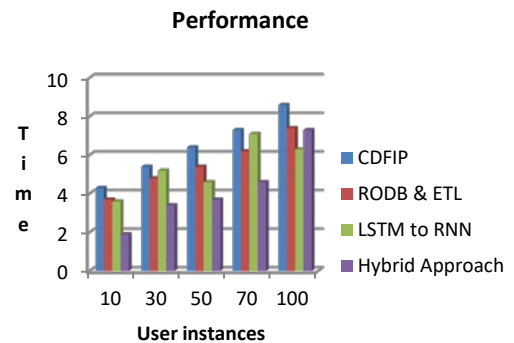


Fig. 1. Performance Graph.

Table III displays a comparison of encryption time for different user instances against Hybrid and existing approaches.

TABLE III. ENCRYPTION TIME COMPARISON

No. of user instances	CDFIP	RODB & ETL	LSTM to RNN	Hybrid Approach
100	4.3	3.7	3.6	3.5
200	5.4	4.8	5.2	4.1
300	6.4	6.7	5.2	4.3
400	7.3	6.2	7.1	4.6
500	8.6	7.4	6.3	7.3

Fig. 2 shows a clear comparison of the reduced Encryption time for proposed in comparison to the existing approaches as user instances keep varying.

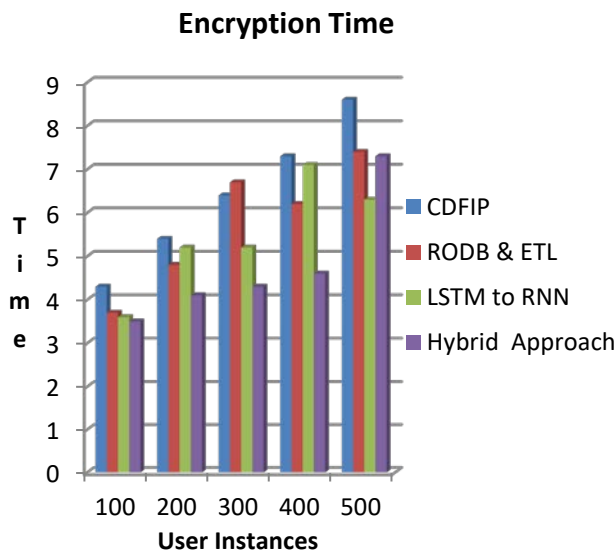


Fig. 2. Encryption Time.

Table IV shows a comparison of decryption time for various user instances against Hybrid and existing approaches.

TABLE IV. DECRYPTION TIME COMPARISON

No of user instances	CDFIP	RODB & ETL	LSTM to RNN	Hybrid Approach
100	3.7	4.7	4.2	3.8
200	4.2	5.6	4.8	3.6
300	3.7	7.4	5.3	4.3
400	6.3	5.82	4.6	4.7
500	5.7	6.4	6.8	5.3

Fig. 3 depicts leveraged decryption throughput of Hybrid approach compared to other approaches.

Table V depicts the memory utilization of different approaches against proposed approach taking into consideration the varying instances of users.

Decryption Time

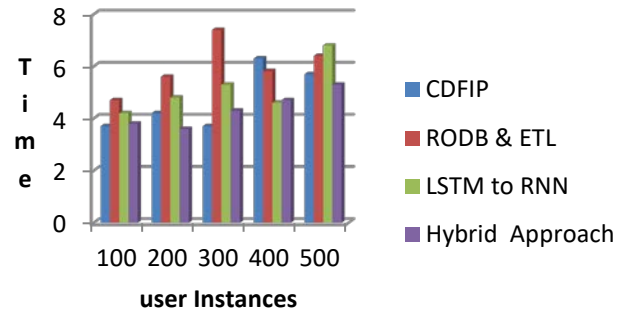


Fig. 3. Decryption Time.

TABLE V. MEMORY UTILIZATION

No of user instances	CDFIP	RODB & ETL	LSTM to RNN	Proposed Approach
100	3541	3642	3876	2759
200	4216	4326	4216	3124
300	6245	5974	3654	3926
400	11021	6785	5243	4146
500	23542	7853	5674	5214

Memory utilization

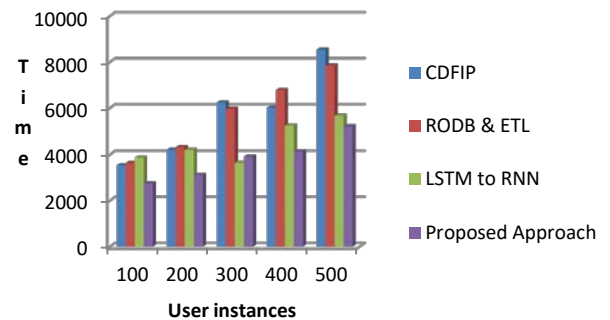


Fig. 4. Graph Showing Memory Utilization.

Fig. 4 shows a clear comparison of the reduced memory utilization by the system in comparison to the existing algorithms.

Hence the simulation results of the proposed algorithm outperform in terms of Performance, Encryption and Decryption throughput and memory utilization providing improved Privacy for patient sensitive data.

VI. CONCLUSION

This paper aimed to discuss the importance of security of patient data based on the access priorities of users, using a Hybrid ABE Approach. In due course several techniques related to mobile healthcare and e-healthcare grabbed concentration in research, but lacked profound architecture to preserve patient data. Our framework offers a innovative and

qualitative technique using SHA-1 and improved Feistel network in key generation ensuring authentication, and confidentiality during transmission entailing limited access to user communities considering access policy. The groups of users with limited access are subjected to Anonymity techniques. The result of our method renders improvised performance in several evaluation parameters considered. Lastly, we conclude that the roadmap presented endeavors a feasible solution for discussed privacy issues.

REFERENCES

- [1] Supriya menon M & Rajarajeswari P, "A Hybrid Machine Learning approach for Drug Repositioning," IJCSNS International Journal of Computer Science and Network Security, VOL.20 No.12, December 2020, <https://doi.org/10.22937/IJCSNS.2020.20.12.24>.
- [2] Saran, P., Rajesh, D., Pamnani, H., Kumar, S., Hemant Sai, T. G., & Shridevi, S, "A Survey on Health Care facilities by Cloud Computing," International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). doi:10.1109/ic-etite47903.2020.231.
- [3] Chen, S.-W., Chiang, D. L., Liu, C.-H., Chen, T.-S., Lai, F., Wang, H., & Wei, W., "Confidentiality Protection of Digital Health Records in Cloud Computing," Journal of Medical Systems, 40(5), 2016, doi:10.1007/s10916-016-0484-7.
- [4] Zhiqiang, G., Lingsong, H., Hang, T., & Cong, L., "A cloud computing based mobile healthcare service system," 2015 IEEE 3rd International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), doi:10.1109/icsima.2015.7559009.
- [5] Haufe, K., Dzombeta, S., & Brandis, K., "Proposal for a Security Management in Cloud Computing for Health Care," The Scientific World Journal, 2014, pp. 1–7, doi:10.1155/2014/146970.
- [6] Singh, I., Kumar, D., & Khatri, S. K., "Improving The Efficiency of E-Healthcare System Based on Cloud," 2019, Amity International Conference on Artificial Intelligence (AICAI), doi:10.1109/aicai.2019.8701387.
- [7] Kumar, J., Goomer, R., & Singh, A. K., "Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters," Procedia Computer Science, 125, 2018, pp. 676–682. doi:10.1016/j.procs.2017.12.087.
- [8] Chiuchisan, I., Balan, D.-G., Geman, O., Chiuchisan, I., & Gordin, I., "A security approach for health care information systems," 2017, E-Health and Bioengineering Conference (EHB). doi:10.1109/ehb.2017.7995525.
- [9] Masood, I., Wang, Y., Daud, A., Aljohani, N. R., & Dawood, H., "Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure," Wireless Communications and Mobile Computing, 2018, 1–23. doi:10.1155/2018/2143897.
- [10] Ronald Glasberg, Michael Hartmann, Michael Draheim, Gerrit Tamm, and Franz Hessel, "Risks and Crises for Healthcare Providers: The Impact of Cloud Computing," 2014, Academic Editors: R. Colomo-Palacios, M. Niedermayer, and V. Stantchev.
- [11] Al-Issa, Y., Ottom, M. A., & Tamrawi, A., "eHealth Cloud Security Challenges: A Survey," Journal of Healthcare Engineering, 2019, 1–15. doi:10.1155/2019/7516035.
- [12] Chenthara, S., Ahmed, K., Wang, H., & Whittaker, F., "Security and Privacy-preserving Challenges of e-Health Solutions in Cloud Computing," IEEE Access, 1–1. doi:10.1109/access.2019.2919982.
- [13] Ramzi A. Haraty, Mirna Zbib and Mehedi Masud, "Data Damage Assessment and Recovery Algorithm from Malicious Attacks in HealthCare Data Sharing Systems," 2016, Secure cloud computing for mobile health services. Peer-to-Peer Networking and Applications, 9(5), 809–811. doi:10.1007/s12083-016-0451-6.
- [14] Azeez, N. A., & der Vyver, C. V., "Security and privacy issues in e-health cloud-based system: A comprehensive content analysis," Egyptian Informatics Journal, 2018, doi:10.1016/j.eij.2018.12.001.
- [15] Anitha, R., & Mukherjee, S., "Data Security in Cloud for Health Care Applications," Advances in Computer Science and Its Applications, 1201–1209, 2014, doi:10.1007/978-3-642-41674-3_167.
- [16] Soceanu, A., Vasylenko, M., Egner, A., & Muntean, T., "Managing the Privacy and Security of eHealth Data," 2015, 20th International Conference on Control Systems and Computer Science. doi:10.1109/cscs.2015.76.
- [17] Awan, I. A., Shiraz, M., Hashmi, M. U., Shaheen, Q., Akhtar, R., & Ditta, A., "Secure Framework Enhancing AES Algorithm in Cloud Computing," Security and Communication Networks, 2020, 1–16. doi:10.1155/2020/8863345.
- [18] Opportunities and Challenges of Cloud Computing to Improve Health Care Services.
- [19] Sun, P. J., "Privacy Protection and Data Security in Cloud Computing: A Survey, Challenges and Solutions," IEEE Access, 1–1. doi:10.1109/access.2019.2946185.
- [20] Supriya menon M & Rajarajeswari P, "A contemporary way for enhanced modeling of context aware privacy system in PPDm," Journal of Advanced Research in Dynamical and Control Systems.
- [21] Uma Narayanan A , Varghese Paul B , Shelbi Joseph A , "A novel system architecture for secure authentication and data sharing in cloud enabled Big Data Environment," Engineering and Technology, Kochi, Kerala India.
- [22] Ma, H., Zhang, R., & Yuan, W., Comments on "Control Cloud Data Access Privilege and Anonymity With Fully Anonymous Attribute-Based Encryption," IEEE Transactions on Information Forensics and Security, 11(4), pp. 866–867, 2016, doi:10.1109/tifs.2015.2509865.

Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features

Oscar Utomo Kumala¹, Amalia Zahra²
Computer Science Department, Bina Nusantara University
Jakarta, Indonesia 11480

Abstract—Emotion recognition is one of the widely studied topics in speech technology. Emotions that come from speech can contain useful information for many purposes. The main aspects in speech emotion recognition are speech features, speech corpus, and machine learning algorithms as the classifier method. In this paper, cross-corpus method is used to conduct Indonesian Speech Emotion Recognition (SER) along with the combination of Mel Frequency Cepstral Coefficients (MFCC) and Teager Energy features. Using Support Vector Machine (SVM) as classifier, the experiment result shows that applying cross-corpus method by adding corpora from other languages to the training dataset improves the emotion classification accuracy by 4.16% on MFCC Statistics feature and 2.09% on Teager-MFCC Statistics feature.

Keywords—Cross corpus; Indonesian speech emotion recognition; Mel Frequency Cepstral Coefficients; Teager Energy

I. INTRODUCTION

Nowadays we are experiencing a rapid growth on Information Technology (IT) sectors, especially in mobile devices area. The interaction between user and mobile devices is getting smoother each day so that it can assist user's daily activities. One important means to achieve this is speech voice. In speech voice there are a lot of information that can be extracted and analyzed. One important aspect of the information is emotion which can contain many additional information, such as the speaker's condition (physical state or mood), the meaning of the speech, and many more.

Different kind of emotion contained in a speech may cause a different response on the person or device the speaker is talking to. A virtual assistant with emotion recognition feature [1] will have advantage of obtaining capability to give different answers or responses depending on the emotion contained in the speech or order. One simple application is the virtual assistant will compile a (song) playlist that is comforting the user if there is sad emotion recognized in the speech.

Because of this high potential of use, it is necessary to further analyze the emotion recognition process itself. From the studies written in literatures, there are three main factors in Speech Emotion Recognition (SER). The first one is speech features, which consist of acoustic features, lexical features, sound volume and frequencies, vocabularies, languages, speaker's background (nationality, ethnic, age, etc.), and many more. The next factor is the availability of corpus which will be used as training and testing set. The last factor is the methods

of machine learning algorithm used to classify the emotion in the speech.

Studies have been carried out at international level for different languages, features [2], corpora [3], methods [4], and algorithms [5] showing various results. Only in recent years, studies for Indonesian have been rising. In this study, we will focus on the Indonesian SER. We need to distinguish SER in Indonesian with other languages because each language and culture has its own characteristic in the SER process [6].

The first main topic in this study is the use of cross-corpus method [7] for the Indonesian SER. We decide to use cross-corpus because of the limitation of the available Indonesian corpus. A SER will achieve better result with larger training dataset. That is the reason we include the corpus from other languages to the training dataset. The corpus used in this study is Berlin Database of Emotional Speech (Berlin EmoDB), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and Surrey Audio-Visual Expressed Emotion (SAVEE). Thus, there are three corpora: one German corpus and two English corpora.

Another main topic is the combination of two speech features, Mel Frequency Cepstral Coefficients (MFCC) features and Teager Energy features. MFCC features are one of the most used speech features in the SER studies [8]. The features will be combined with Teager Energy features [9] to hopefully achieve better result. These speech features are extracted from the corpus and used along with their statistical values. The combination of the features will be tested together with the cross-corpus method and Support Vector Machine (SVM) classifier. The results of the test will be analyzed thus a number of conclusions can be drawn.

The remainder of this paper is as follows. In Section II, we discuss the previous research and studies that are relevant to this topic. We describe the configurations of our experiments in Section III. In Section IV, we describe the results of our experiments and perform analysis on them. Section V concludes the paper.

II. RELATED WORKS

Studies on SER started at international level. One of the first studies was conducted by Shah & Hewlett [10] to detect emotion by extracting and analyzing speech features which consist of pitch, MFCC, and Formants. The study used SVM

with Linear kernel and k-Mean as classifiers. It showed that the emotion recognition rate is higher for male speakers compared to female ones, and there are some similar emotions: happy, elation, and interest; agitated and subdued.

A framework was created by Pfister [11] to detect emotion by analyzing speech in real time, extracting speech features using OpenSMILE algorithm: energy, volume, voice quality, mel-spectra, MFCC, and some calculations such as mean, extreme, peak, percentile, and deviation. Using SVM with Radial Basis (RBF) kernel, it achieved 70% to 89% accuracy rate depending on the selected method with low delay (0.046 to 0.110 second for 1- to 5-second sentence). The notable finding from the framework is that emotion recognition can be performed in real time, which is vital to human-computer interaction applications.

Studies on Indonesian SER have started in recent years [12]. There is a study by Lubis, Lestari, Purwarianti, Sakti, & Nakamura [13] which succeeded in forming the first Indonesian emotional speech corpus, namely Indonesian Emotional Speech Corpus (IDESC). From the study, it can be concluded that the recognition of angry emotion has a relatively higher accuracy than satisfied emotion. In general, active emotions are easier to recognize than passive emotions. The study achieved 68.31% accuracy for the classification of four emotion classes. Another study presents a speaker-independent emotion recognition [14]. The study also found that disgust is the most difficult emotion to detect, followed by sad emotion.

One of the latest Indonesian SER studies was conducted by Gunawan & Idananta [15] by testing sound signals in Indonesian. Existing sound signals are analyzed using the MFCC features combined with the Teager Energy feature. The emotions were classified using SVM into four classes, namely angry, fear, happy, and sad. The speech corpus was created by recording conversations from four amateur actors and actresses. They speak 15 Indonesian sentences for four times, each based on the emotion requested. This speech corpus will also be used in this study as Indonesian corpus.

The results in [15] show that Teager Energy is an important feature that contributes to the accuracy of emotional classification by approximately 41%. The study also found that happy emotion seemed to be somewhat difficult to distinguish from angry emotion. In addition, the emotions of angry, fear, and sad can be recognized from speech signals with high accuracy.

III. EXPERIMENTS

We begin the experiments by preparing the training and testing dataset which consist of the corpora aforementioned. After the training and testing dataset are ready, the speech feature extraction process starts, followed by training and testing process, which is followed by evaluation.

A. Preparing Corpus

There are four corpora used as training and testing dataset. The first corpus is Berlin EmoDB [16], which is a database of German emotional speech containing 535 audio files with 7 emotion classes. The second one is RAVDESS [17], which is a

validated multimodal database of English emotional speech and song with North American accent, containing 1440 audio files with 8 emotion classes. The next corpus, SAVEE [18], is an English audiovisual database with British accent which consists of 480 audio files with 7 emotion classes. The Indonesian corpus is the last corpus used with 4 emotion classes, resulting in 60 audio files.

All the corpora will go through emotion filter process. The experiment will only use 4 emotion classes: angry, fear, happy, and sad. All emotion classes outside those 4 will not be included in the training and testing dataset. After this process, there is 61% data left on Berlin EmoDB, 54% data left on RAVDESS, 50% data left on SAVEE, and 100% data on Indonesian corpus.

The next process is data standardization process, which balances the data amount of all emotion classes. This process is necessary because the data amount of each emotion class is not the same for each corpus. Besides standardizing the data amount, the audio bitrate of all audio files is also standardized into 256 kbps with the assistance of Audacity desktop application.

Through emotion filter process and data standardization processes, the data amount of each class emotion becomes 61 on Berlin EmoDB, 174 on RAVDESS, 60 on SAVEE (same as original), and 60 on Indonesia corpus (same as original). There are two data excluded from RAVDESS because we are unable to extract Teager Energy feature from the audio files. The final amount of data that can be used for training and testing dataset is 1418.

After the previous processes, all corpora will be combined and then divided into three corpus groups: 100% corpus group, 80% corpus group, and 20% corpus group. 100% corpus group is the fully combined corpus. 80% corpus group consists of 80% data of combined corpus, where the data are picked manually, and used as part of training dataset. Likewise, 20% corpus group consists of 20% data of combined corpus, where the data are picked manually, and used as part of testing dataset. By performing such a grouping, the ratio of 80% and 20% for each corpus can be achieved.

B. Extracting the Speech Features

There are two kinds of speech features that will be used for training and testing process: MFCC features and Teager Energy features. We begin the speech feature extraction process by reading the corresponding audio file with wav file extension. Reading the audio file will give us the signal and rate values which will be used to calculate the MFCC and Teager Energy values.

The MFCC feature extraction process is carried out using `python_speech_features` library [19]. A simple function in the library takes the signal and rate values as parameters and return the MFCC feature values in the form of a 2-dimensional array. The size of the array is 13 times the number of sound frames. The number 13 is obtained from the number of frequency bands in a speech voice, and the number of sound frames produced depending on the duration of the corresponding audio file. In this study, the number of sound frames taken is 75, which is the smallest number of sound frames from all corpus

data. Thus, we will get 975 MFCC feature values saved into a database in the form of a 1-dimensional array.

In addition to the MFCC feature values, statistical values will also be calculated for each frequency band. These statistical values include mean (average array value), min (smallest array value), max (largest array value), std (array standard deviation), and median (array middle value). Thus, 5 statistical values will be obtained for each of the 13 frequency bands so that a total of 65 MFCC statistical values will be saved into the database in the form of a 1-dimensional array.

Similar to the MFCC feature extraction process aforementioned, the Teager Energy feature extraction process is also carried out using a library. The library has a function to return the values of Teager-Kaiser Operator (also known as the Nonlinear Energy Operator) and Envelope Derivative Operator (EDO). The Teager-Kaiser Operator values will be used in this study as the Teager Energy feature values. The values are returned in 1-dimensional array with the size depending on the duration of the corresponding speech file. In this study, the first 1000 arrays will be retrieved and divided into two features (500 arrays each), namely Teager 1 feature and Teager 2 feature. Both features are saved to the database in the form of a 1-dimensional array.

In addition to the Teager Energy feature values, the statistical values are also calculated. Similar to the MFCC statistical values, Teager Energy statistical values also include mean (average array value), min (smallest array value), max (largest array value), std (array standard deviation), and median (array middle value). Thus, we will obtain five Teager Energy statistical values saved to the database in the form of a 1-dimensional array.

C. Configuring Corpus for Training and Testing

Training and testing processes are interconnected. A training process produces a model used for one or more testing processes. Both processes are conducted to as many corpus combinations as possible so we can obtain many testing results for this study.

From the initial four corpora, there will be two additional corpora formed from the combination of those corpora. The first one is International corpus which consists of three corpora in non-Indonesian language: Berlin EmoDB, RAVDESS, and SAVEE. Another one is a combined corpus which consists of all initial four corpora. Thus, there are six corpora in total for training and testing processes.

With such corpus combinations, we need a proper configuration for the use of corpus grouping in certain training and testing scenarios. Table I shows the possible scenarios and the configuration.

First scenario is when the training and testing processes use the same corpus. Here we will use 80% corpus group for training and 20% corpus group for testing. The next scenario is the opposite of first scenario, when the training and testing processes use the different corpora. We will use 100% corpus group for both processes.

The third scenario is when the training process uses a combined corpus which consists of a single corpus that is also

used for testing process. In this case, we will use 80% corpus group at the training process and 20% corpus group at the testing process. The other corpus of combined corpus at the training process will use 100% corpus group.

The final scenario is the opposite of the third scenario, when training process uses a single corpus which is also used as part of a combined corpus at testing process. In this case, for that single corpus we will use 80% corpus group at the training process and 20% corpus group at the testing process. The other corpus of combined corpus at the testing process will use 100% corpus group.

D. Conducting Training and Testing

In this study, we conduct the training using SVM with RBF kernel. First step of the training process is retrieving speech features values. There are five speech features: MFCC feature, MFCC Statistics feature, Teager Energy 1 feature, Teager Energy 2 feature, and Teager Energy Statistics feature.

At this point, we add a new speech feature which is formed from the combination of Teager Energy Statistics feature with MFCC Statistics feature, namely Teager-MFCC Statistics feature. Thus, there are six speech features values which can be used for next processes.

All speech feature values obtained need to go through normalization process. This normalization process is called scaling in Python, where all values will be normalized to the range of -1 to 1. Normalization will be carried out on the training speech feature values so that the minimum value is -1 and the maximum value is 1. The normalization will produce a scale which will be applied to the testing speech features values.

The next process is building model. This process is carried out by importing 'RandomizedSearchCV' from 'sklearn.model_selection' in Python. This module aims to form the best model by looking for random combinations of parameter values from several predetermined parameter values. The mentioned parameters are C and Gamma values. In this module, 10-fold cross validation is applied to the training dataset. The 'RandomizedSearchCV' process is repeated 100, 250, and 500 times. The next process is testing. Each model built from the previous step will be tested using the testing dataset to classify emotions.

TABLE I. CONFIGURATION FOR THE USE OF CORPUS GROUPING IN CERTAIN TRAINING AND TESTING SCENARIOS. # IS NUMBER OF SCENARIO, TRC IS TRAINING CORPUS, TSC IS TESTING CORPUS, TRG IS TRAINING CORPUS GROUPING, TSG IS TESTING CORPUS GROUP

#	TRC	TSC	TRG	TSG
1	A	A	80% A	20% A
2	A	B	100% A	100% B
3	A B C	A	80% A 100% B 100% C	20% A
4	A	A B C	80% A	20% A 100% B 100% C

IV. RESULTS AND DISCUSSION

We divide the testing results into three parts. The first part is the testing results for the same corpus. The next part is the testing results for different corpus. The final part is the analysis of all testing results.

A. Testing Result for the Same Corpus

Table II shows the accuracies of testing using the same corpus. All six available corpora will go through testing process with six speech features values. The results shown in the table are those that achieve the best average accuracy among three numbers of iteration (i.e. 100, 250, and 500). The configuration of corpus grouping is applied here.

B. Testing Result for different Corpus

In contrast with testing using the same corpus in Table II, The result of testing using different corpus is shown in Table III. All six available corpora will go through testing process with several corpus combinations and five speech features values. The Teager Energy 2 feature is excluded here because the result is very similar to that using the Teager Energy 1 feature. The results shown in the table are those that achieve the best average accuracy between three numbers of iteration (i.e. 100, 250, and 500). The configuration of corpus grouping is also applied here.

TABLE II. TESTING RESULT FOR SAME CORPUS. C IS CORPUS NAME, F1 IS MFCC FEATURE, F2 IS MFCC STATISTICS FEATURE, F3 IS TEAGER 1 FEATURE, F4 IS TEAGER 2 FEATURE, F5 IS TEAGER STATISTICS FEATURE, F6 IS TEAGER-MFCC STATISTICS FEATURE, AVG IS AVERAGE ACCURACY, C1 IS BERLIN EMODB, C2 IS RAUDESS, C3 IS SAVEE, C4 IS INDONESIAN, C5 IS INTERNATIONAL, C6 IS COMBINED CORPUS

C	F1	F2	F3	F4	F5	F6	Avg
C1	68.75%	87.50%	47.92%	41.67%	47.92%	85.42%	63.20%
C2	41.18%	83.82%	25.74%	20.59%	45.59%	79.41%	49.39%
C3	54.17%	58.33%	39.58%	39.58%	31.25%	58.33%	46.87%
C4	54.17%	79.17%	43.75%	47.92%	66.67%	83.33%	62.50%
C5	45.69%	80.17%	29.31%	30.17%	39.66%	81.90%	51.15%
C6	46.43%	82.50%	25.71%	26.43%	40.00%	80.36%	50.24%
Avg	51.73%	78.58%	35.34%	34.39%	45.18%	78.13%	53.89%

TABLE III. TESTING RESULT FOR DIFFERENT CORPUS. TR IS TRAINING DATASET, TS IS TESTING DATASET, F1 IS MFCC FEATURE, F2 IS MFCC STATISTICS FEATURE, F3 IS TEAGER 1 FEATURE, F5 IS TEAGER STATISTICS FEATURE, F6 IS TEAGER-MFCC STATISTICS FEATURE, AVG IS AVERAGE ACCURACY, C1 IS BERLIN EMODB, C2 IS RAUDESS, C3 IS SAVEE, C4 IS INDONESIAN, C5 IS INTERNATIONAL, C6 IS COMBINED CORPUS

TR	TS	F1	F2	F3	F5	F6	Avg
C1	C2	25.50%	28.53%	25.07%	31.84%	26.51%	27.06%
C1	C3	35.83%	36.25%	29.17%	26.67%	22.08%	31.17%
C1	C4	42.50%	49.17%	17.08%	23.33%	37.92%	29.83%
C1	C5	29.84%	34.11%	26.88%	36.25%	33.20%	30.53%
C1	C6	30.36%	29.54%	24.55%	35.02%	28.81%	28.69%
C2	C1	25.00%	50.82%	25.41%	24.59%	47.13%	30.41%
C2	C3	23.33%	26.25%	25.00%	24.17%	27.92%	24.75%
C2	C4	23.33%	29.58%	25.00%	27.08%	32.92%	26.00%
C2	C5	27.10%	46.94%	24.52%	25.65%	44.52%	29.97%
C2	C6	26.51%	42.33%	25.12%	29.65%	41.63%	29.84%
C3	C1	25.00%	39.75%	42.21%	16.80%	32.79%	32.87%
C3	C2	22.19%	25.36%	24.93%	29.11%	25.36%	25.30%
C3	C4	25.83%	26.25%	14.17%	20.42%	26.25%	20.33%
C3	C5	27.28%	30.93%	30.43%	26.17%	32.45%	28.88%
C3	C6	26.35%	35.73%	27.49%	25.69%	33.93%	28.39%
C5	C1	66.67%	89.58%	45.83%	54.17%	89.58%	59.17%
C5	C2	34.56%	84.56%	24.26%	43.38%	84.56%	42.20%
C5	C3	43.75%	75.00%	37.50%	27.08%	79.17%	43.33%
C5	C4	34.17%	48.33%	16.25%	25.83%	44.58%	28.25%
C6	C4	58.33%	83.33%	35.42%	43.75%	85.42%	46.25%
C6	C5	44.40%	79.74%	26.72%	37.07%	74.14%	43.10%
Avg		33.36%	47.24%	27.48%	29.93%	45.28%	34.81%

C. Analysis

Analysis is conducted to the results obtained from both testing using the same corpus and that using different corpus. From the result of testing using the same corpus, we can highlight several points. The first point is related to speech features that achieve the highest accuracy for each corpus. MFCC Statistics feature achieves the highest accuracy on Berlin EmoDB (87.50%), RAVDESS (83.82%), SAVEE (58.33%, tied with Teager-MFCC Statistics feature), and Combined corpus (82.50%). Meanwhile, Teager-MFCC Statistics feature achieves the highest result on SAVEE, Indonesian corpus (83.33%), and International corpus (81.90%).

Next point is that Berlin EmoDB has the highest average accuracy among all corpora (63.20%), while SAVEE has the lowest one (46.87%). The last point to highlight is that MFCC Statistics feature achieves the highest average accuracy (78.58%) followed tightly (0.45%) by Teager-MFCC Statistics feature (78.13%). On the contrary, Teager Energy 2 feature has the lowest average accuracy (34.39%) followed by Teager Energy 1 feature (35.34%).

We can see different highlights on the testing result for different corpus. Combined corpus achieves the highest average accuracy (44.68%) followed tightly (1.44%) by International corpus (43.24%). From the result, it is shown that corpus which consists of many single corpora (International corpus and Combined corpus) has higher average accuracy than single corpus (Berlin EmoDB, RAVDESS, SAVEE, and Indonesian corpus). Additionally, similar to the result of testing using the same corpus, MFCC Statistics feature achieves the highest average accuracy (47.24%) followed tightly (1.96%) by Teager-MFCC Statistics feature (45.28%). Teager Energy 2 feature once again has the lowest average accuracy (25.57%) followed by Teager Energy 1 feature (27.48%).

From both testing results, we can highlight several points. The order of the speech features that achieve from the highest to the lowest average accuracy is MFCC Statistics feature, Teager-MFCC Statistics feature, Teager Statistics feature, MFCC feature, Teager Energy 1 feature, and Teager Energy 2 feature. Even though MFCC Statistics feature achieves the highest average accuracy, in some cases Teager-MFCC Statistics produces a better result.

Another point to highlight is that statistical features (MFCC Statistics, Teager Statistics, and Teager-MFCC Statistics) achieve higher average accuracy than non-statistical features (MFCC, Teager Energy 1, and Teager Energy 2). The average accuracy of testing using the same corpus is still higher than that using different corpora.

For specific case where Indonesian corpus used as a testing dataset, we obtained a higher accuracy rate when using combined corpus as the training dataset compared to that achieved by using the same corpus. We achieved the accuracy of 83.33% and 79.17% from testing using the MFCC Statistics feature for the first and latter scenario, respectively, whereas using Teager-MFCC Statistics feature achieved the accuracy of 85.42% and 83.33% for such scenarios, respectively. It means that there is an accuracy improvement by 4.16% using MFCC

Statistics feature and 2.09% using Teager-MFCC Statistics feature.

V. CONCLUSION

Based on the results and highlights described in the previous section, we can conclude three main points. First, we can see that applying cross-corpus method by adding corpora from other languages to the training dataset can improve the overall performance of the emotion recognition, including the Indonesian SER. From the corpora, we can also see that English and German languages have good compatibility with Indonesian in emotion classification aspect.

Next, we can see that when applying cross-corpus method in Indonesian SER, the speech features that achieve the best results are MFCC Statistics feature and Teager-MFCC Statistics feature. And finally, the accuracy improvement of 4.16% using MFCC Statistics feature and 2.09% using Teager-MFCC Statistics feature could be a good start for Indonesian SER and can be further improved in the future.

Potential improvement in future studies may include the use of more complex speech features complemented with feature selection method and other classification methods apart from SVM. A good classifier that is worth a try is Extreme Learning Machine (ELM) [20] which has proven to achieve good results in some studies.

REFERENCES

- [1] G. Iannizzotto, L. Lo Bello, A. Nucita, and G. M. Grasso, "A Vision and Speech Enabled, Customizable, Virtual Assistant for Smart Environments," Proc. - 2018 11th Int. Conf. Hum. Syst. Interact. HSI 2018, no. November, pp. 50–56, 2018.
- [2] T. Özseven, "A Novel Feature Selection Method for Speech Emotion Recognition," Appl. Acoust., vol. 146, pp. 320–326, 2019.
- [3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011.
- [4] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition Using Fourier Parameters," IEEE Trans. Affect. Comput., vol. 6, no. 1, pp. 69–75, 2015.
- [5] L. Yu, B. Wu, and T. Gong, "A Hierarchical Support Vector Machine Based on Feature-driven Method for Speech Emotion Recognition," pp. 901–908, 2013.
- [6] N. Kamaruddin, A. Wahab, and C. Quek, Cultural Dependency Analysis For Understanding Speech Emotion, Expert Syst. Appl., vol. 39, no. 5, pp. 5115–5133, 2012.
- [7] H. Kaya and A. A. Karpov, "Neurocomputing Efficient and effective strategies for cross-corpus acoustic emotion recognition," Neurocomputing, vol. 275, pp. 1028–1034, 2018.
- [8] C. S. Ooi, K. P. Seng, L. M. Ang, and L. W. Chew, A new approach of audio emotion recognition, Expert Syst. Appl., vol. 41, no. 13, pp. 5858–5869, 2014.
- [9] L. Kerkeni, Y. Serrestou, K. Raouf, M. Mbarki, M. A. Mahjoub, and C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO, Speech Commun., vol. 114, no. May, pp. 22–35, 2019.
- [10] R. Shah and M. Hewlett, "Emotion detection from speech, Final Proj. cs, p. 229, 2007.
- [11] T. Pfister, Emotion Detection from Speech, Gov. Caius Coll., 2010.
- [12] F. Kasyidi and D. P. Lestari, "Identification of Four Class Emotion from Indonesian Spoken Language Using Acoustic and Lexical Features," J. Phys. Conf. Ser., vol. 971, no. 1, 2018.
- [13] N. Lubis, D. Lestari, A. Purwarianti, S. Sakti, and S. Nakamura, Emotion recognition on Indonesian television talk shows, pp. 466–471, 2014.

- [14] M. Kurniawati, Pipin; Lestari, Dessi Puji; Leylia Khodra, Speech emotion recognition From Indonesian spoken language using acoustic and lexical features, no. November, pp. 1–3, 2017.
- [15] F. E. Gunawan and K. Idananta, Predicting the level of emotion by means of Indonesian speech signal, TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 15, no. 2, p. 665, 2018.
- [16] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, A database of German emotional speech, 9th Eur. Conf. Speech Commun. Technol., no. January, pp. 1517–1520, 2005.
- [17] S. R. Livingstone and F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, vol. 13, no. 5. 2018.
- [18] P. Jackson and S. ul haq, Surrey Audio-Visual Expressed Emotion (SAVEE) database, 2011. [Online]. Available: <http://kahlan.eps.surrey.ac.uk/savee/>. [Accessed: 26-May-2020].
- [19] James Lyons et al., “jameslyons/python_speech_features: release v0.6.1,” 2020.
- [20] W.-H. Cao, M. Wu, J.-P. Xu, Z.-T. Liu, G.-Z. Tan, and J.-W. Mao, “Speech Emotion Recognition Based on Feature Selection and Extreme Learning Machine Decision Tree,” *Neurocomputing*, vol. 273, pp. 271–280, 2017.

NetAI-Gym: Customized Environment for Network to Evaluate Agent Algorithm using Reinforcement Learning in Open-AI Gym Platform

Varshini Vidyadhar¹

Research Scholar, Department of
Computer Science and Engineering
Bangalore Institute of Technology
Bangalore, India

Dr. Nagaraj R²

Professor, Department of
Information Science and
Engineering, Bangalore Institute of
Technology, Bangalore, India

Dr. D V Ashoka³

Professor, Department of
Information Science and
Engineering, JSS Academy
Technical Education, Bangalore, India

Abstract—The growing size of the network imposes computational overhead during network route establishment using conventional approaches of the routing protocol. The alternate approach in contrast to the route table updating mechanism is the rule-based method, but this also provides a limited scope in the dynamic networks. Therefore, reinforcement learning promises a better way of finding the route, but it requires an evaluation platform to build a model synchronization between route and agent. Unfortunately, the de-facto platform for agent evaluation, namely Open-AI Gym, does not provide a suitable networking environment. Therefore, this paper aims to propose a networking environment as a novel contribution by designing a suitable customized environment for a network synchronically with Open-AI Gym. The successful deployment of the proposed network environment: NetAI-Gym provides a functional and practical result that can be used further to develop routing mechanisms based on Q-learning. The validation of the proposed NetAI-Gym is carried out with different nodes in the network regarding Episodes Vs. Reward. The experimental outcome justifies the validity of the proposed NetAI-Gym that it is suitable for solving network-related problems.

Keywords—Open-AI Gym; network; environment; agent; reinforcement learning

I. INTRODUCTION

Artificial intelligence (AI) is being explored way back in 1997 for some problems like exploring the possibility of adaptive-AI using a network of neurons like adaptive elements, where the focus of the study was on the adaptive systems, where the learning system adapts some behavior from the environment to maximize the signal. It is being observed that this approach has received very little attention from the researchers from the computational perspectives [1][2]. At the same time, the same idea of the hedonistic-learning system (HLS) of that time has been realized today as Reinforcement Learning (RL). However, with a hypothesis that data are collected only from the IEEE Digital library. It is found that the routing problem in the network became an active research problem in the last 20 years, with an overall publication of 86,226. It is observed that in the last decade, the total publication for the same problem is 52,344, which is alone 60.7%, which shows that the active focus of the researchers is higher in the running decade. Considering this 60.7% data as

100% and then the stake of Reinforcement Learning is found only 322 in totality, which is hardly 0.6% and 0.3% from last two decade. Therefore, it can be concluded that more efforts are required to study and develop a solution paradigm for routing problems in a network using reinforcement learning. The typical architecture of reinforcement learning is shown in Fig. 1.

The basic design of RL includes building two functional blocks, namely E and Ag. The Ag takes appropriate Ac based on the O provided by E and subsequently based on the Ac taken, E gives positive or negative R. Therefore, to evaluate the agent algorithm, a suitable platform of the environment is required as per the domain context and particular task. The role of RL is to solve the problem of sequential decision tasks in different networking scenarios. There are many methods found in the literature to solve this problem by using i) Game theory [3], ii) Swarm [4], iii) a probabilistic technique [5], and many more [6-7]. However, all these approaches are associated with some advantages and limitations. But RL can be utilized to address very complex problems that conventional approaches cannot address. RL refers to the computer intelligence field that studies programmed computing procedures and dynamically optimizes their performance based on experience learned from the environment. Therefore, RL offers promising context that can be used to develop adaptive mechanisms for network routing so that better performance can be achieved on complex problems without performing any engineering particular to the problem. RL's logic considers a decision-maker component (agent) in the environment (set of states with inputs). At every step, the agent takes action and gets observations and rewards when interacting with the environment. The RL algorithm tries to maximize a certain amount of reward achieved by the agent. The RL environment for networking was configured based on a general backbone network according to the concept of a partially observable Markov decision process [8]. However, most of the researchers failed to produce their experiments based on the RL. Recently, an introduced RL tool kit, namely Open-AI Gym, removes this problem and lacking standardization in the research process by giving versatile numbers of the environment with great ease of setting up. This toolkit offers a collection of test problems. It concentrates on RL's scenario setting, in which the experience learned by the

agent is divided into several episodes. The Open-AI provides a benchmarking framework for building and testing RL algorithms. However, to date, no any Open-AI Gym library is available for networking. A thorough investigation of the existing Open-AI Gym platform reveals that it has 7 explicit classes contributed by Open-AIGym and an additional class of third-party contribution. A detailed explanation is given in section III. It is found that the available environment in Open-AI Gym is not applicable for solving the problem of network, especially the routing. Researchers use various network simulators and experimental testbeds. Hence, this paper is the first of its type to contribute a custom design of an environment for evaluating network routing using RL on the Open-AI Gym platform. The proposed NetAI-Gym offers a scalable networking environment for implementing any reinforcement learning algorithm for training agents and accessing their performances in the context of networking. This paper is organized as in Fig. 2.

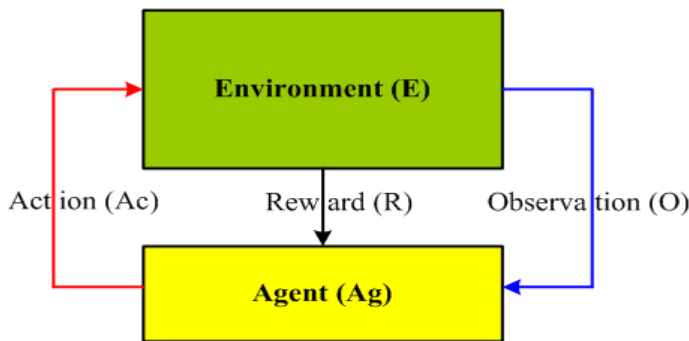


Fig. 1. Typical Reinforcement Learning Architecture.

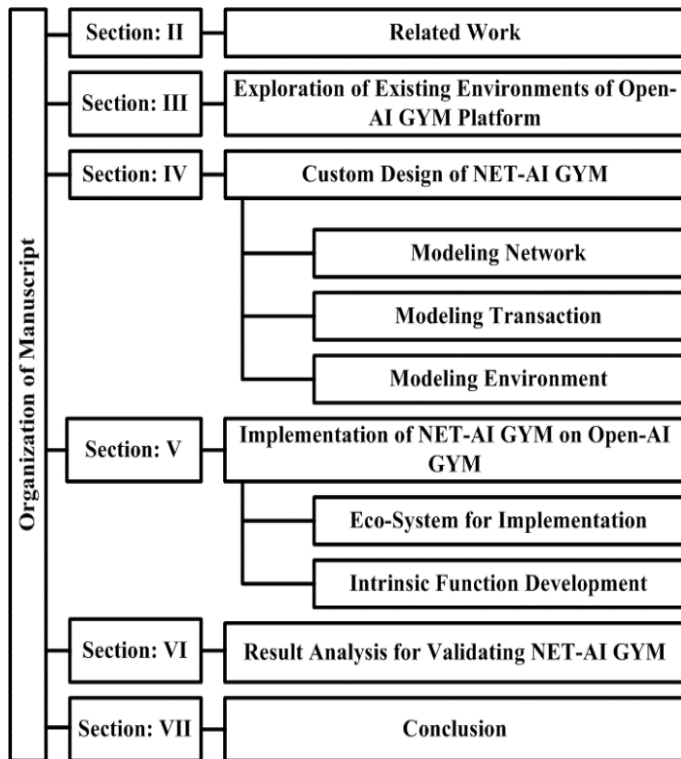


Fig. 2. Organization of the Paper.

II. RELATED WORK

The use of RL is found in the literature for the various task in the respective networks. In WSN, automation of the radio scheduling task for optimizing energy usage is designed using the RL technique [9]. RL eliminates overheads of the control messages used for communication among neighbors as \forall nodes \in WSN as it approximates its neighbor conditions based on the current state. Since this method does not use any specific tool like Open-AI Gym to evaluate the agent designed, it uses an approach of trial and error to simplify the problem of the sequential decision using game theory. However, it is a computationally expensive method. Reinforcement learning aims to solve sequential decision tasks through trial-and-error interactions with the environment. Another related network in the context of industrial-WSN uses RL for minimizing the latency and maximizing the lifetime of the network [10]. There remain many open issues that include computation of time complexities in terms of learning and exploration that validate performance enhancements. A Markov decision process is used to formulate a path selection process. A deep RL technique is then applied to minimize the probability of network congestion in the network under heavy traffic [11]. Throughput is achieved, but overall network performance can be enhanced by considering network-related dynamic parameters. RL technique adoption is also found in [12] for designing grid-oriented routing mechanisms to address message forwarding challenges in the Vehicular ad-hoc network (VANET). However, the focus is only on the problem of the message forwarding issue from the source to the fixed destination. The Q-table is learned offline and may not be well suited to the dynamic characteristics of urban VANETs. Traffic-aware and road-side-unit (RSU) supported routing mechanism is introduced in [13]. Here, RL is implemented to facilitates intelligent data transmission processes between vehicle-to-vehicle and RSU-2-RSU. However, the vehicle's direction is not considered, which may affect the performance of the routing scheme when it comes to real-time deployment scenarios. An RL-based routing protocol is designed in [14] to analyze the impact of a varying number of nodes on the performance of the Underwater Acoustic Sensor Networks (UWSN). Q-learning is used, where the node has packets to forward based on the state of the buffer, remaining energy, and proximity of adjacent nodes to select the next sender node. In [15], a channel-aware RL-oriented adaptive path selection technique is introduced for multi-hop UWSN. The protocol switches between single-path and multi-path routing accomplish joint optimization in energy consumption and packet delivery ratio. Q-learning-based distributed opportunistic routing mechanism is introduced by [16] for minimizing the average packet routing cost in Wireless Ad-hoc network. This mechanism combinedly solves the problems related to learning and routing network structure is characterized by the communication and data transaction success probability. An efficient mechanism for collaborative RL is used in [17] for optimizing path selection in MANET. The use of RL is used for routing optimization in software-defined networking (SDN) [18]. The effectiveness of the agent is tested under the self-convergence aspects. The authors in [19] used a deep RL mechanism for optimizing routing performance in SDN. The authors in [20] explored RL's

effectiveness towards the energy harvesting routing model based on Q-learning for multi-hop Cognitive Radio networks (CRN). The runtime complexity of this model is $O(N^2)$. However, the performance of CRN can be further enhanced using spectrum sensing and power allocation mechanism. The problem of link selection in the Energy Harvesting Relay network (EHRN) is solved by [21] using RL and Deep-Q-learning techniques. A pre-trained algorithm is used to avoid the massive iterations and alleviate the computation overhead in convergence optimization. However, this approach is not much scalable when environmental parameters change. The use of RL in [22] is found for designing routing protocol in Magnetic Induction Underwater Sensor Networks (MIUSN). A Q-table is derived by taking into account the distance factor and energy loss. However, the protocol requires periodic control message exchange for neighbor discovery to give rise to high overheads and reduce channel usage due to slow propagation. The adoption of the RL-based Q-learning technique for content placement is found in [23] for a dynamic cloud content delivery network (CCDN). An efficient routing design based on RL in Unmanned Robotic Network (URN) is suggested by [24] considering location information, link condition, and battery information to realized the neighbor node with the most significant future reward for determining the next hop. Table I summarizes the above-discussed literature for quick insight concerning network scenarios and issues handled.

It is analyzed that RL is mainly adopted for addressing optimization problems of routing and congestion control in various networking scenarios. However, none of the studies are found in the context of using a customizable environment for the agent algorithm designed to solve routing and network performance problems using RL. The environment used in the existing literature is formulated based on the general core network in the simulation process. Since RL is associated with reasonable overhead, the existing approaches are not suitable for providing an efficient solution. It may be exposed to many problems when it comes to real-time deployment scenarios. Therefore, a customizable environment specific to RL approaches for networking is required for suitable analysis of performance.

TABLE I. SUMMARY OF EXISTING LITERATURE

Authors	Network Scenario	Problem Handled
5, 6	WSN	Energy usage [5], Latency [6]
7	Relay	Network congestion
8,9	VANET	Message forwarding [8], transmission efficiency [9]
10,11	UWSN	Energy
12	Ad-Hoc	Routing cost
13	MANET	Path selection
14,15	SDN	Routing Performance
16	CRN	Network Performance
17	EHRN	Computation overhead
18	MIUSN	Energy and Channel utilization
19	CCDN	Content placement
20	URN	Routing and Energy

III. EXPLORATION OF EXISTING ENVIRONMENTS OF OPEN-AI GYM PLATFORM

The first instance of Open-AI Gym is found in 2016 to support reinforcement learning used for various decision making and control systems. It includes a study of agent learning to achieve the learning goal in an uncertain and complex environment. The use of RL is found in the diversified problem domain, such as robot motor control, games, and business decision makings, wherever a sequence of decision making is required. In the recent past, RL is being used in various complex environments. However, the advancement into deep learning demands engineering aspects specific to the problem.

A. Custom Design of Net-Ai Gym

1) *Modelling network*: The mathematical model for the network is represented by a collection of vectors as in set: $\eta = \{\vec{N}_1, \vec{N}_2, \vec{N}_3, \dots, \vec{N}_n\}$, where, $\forall \vec{N}_k \in \eta$ represents a node with two intrinsic properties as $\{\text{Node number } (X_k), \text{Set of links } (R_k)\}$ s.t \vec{N}_k is represented by a pair of $\{X_k, R_k\}$, where $k = 1$ to n , $n \in \mathbb{N}$ and $n \geq 2$. Basically, a node \vec{N}_k may have connectivity with many of the node $\in \eta - \{\vec{N}_k\}$, therefore, \forall the link-set is represented by a collection of vectors: $R_1 = \{\vec{L}_1, \vec{L}_2, \vec{L}_3, \dots, \vec{L}_m\}$ s.t $\forall \vec{L}_k \in R_1$ contains two intrinsic properties as $\{\text{Connecting node number } (X_k), \text{the weight of the link } (W_k)\}$. This arrangement of vector representation for the nodes and reference to the links alleviates the challenge of handling memory usage by dimensionality reduction. Else the simple representation of Tensor imposes excessive use of memory and computational resources. Fig. 3 illustrates a sample representation of a weighted 3 node network with 3 links.

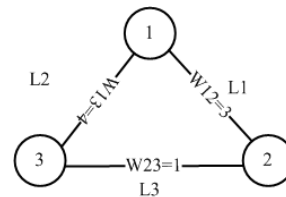


Fig. 3. Network Representation.

$\eta = \{\{X_1, R_1\}, \{X_1, R_2\}, \{X_1, R_3\}\}$, where, $R_1 \rightarrow \{\{X_2, 3\}, \{X_3, 4\}\}$, $R_2 \rightarrow \{\{X_1, 3\}, \{X_3, 1\}\}$, $R_3 \rightarrow \{\{X_1, 4\}, \{X_2, 1\}\}$

2) *Modeling transaction*: The transaction (T) is the delivery of one packet (P) from the source node (Ns) to the destination node (Nd). There are two outcomes of the transaction: i) FAILURE and ii) SUCCESS. To elaborate the process of the transaction and associate states, Fig. 4 with 4-nodes considering node N1 as Ns and node N4 as Nd.

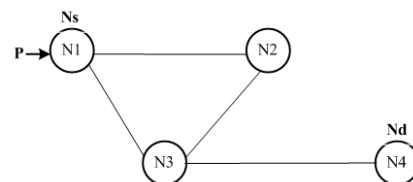


Fig. 4. Illustration of Transaction.

There are various possibilities of transactions between N1 to N4. The first possible condition is that when node N1 does not find node N4 in its first transaction, then it DROPS the packet P, and the state is flagged to FAILURE. Based on the previous record of T, in the next transaction, N1 will look for another node, say node N2, and this time also the state is FAILURE as when a packet is delivered to N2, no connection is found between node N2 and N4. Therefore, in reference to Fig. 3, the only possible transaction for SUCCESS state is to transfer packet P either from N1 (Ns) → N2→ N3→N4 (Nd) or from N1 (Ns) → N3→N4 (Nd) and in this way packet is DELIVERED from Ns to Nd. The computation process for one transfer where a transfer is a process of forwarding a packet from one node to another node is as in algorithm 1.

```

Algorithm-1: Computational process of transfer T(Nk, Nk+1)
Input: η
Output: REWARD, DONE
Start
1. Initialize Nk, Nk+1
2. T (Nk→Nk+1 )
3. GET Rk∈ Nk
4. GET Xk+1 ∈ Rk
5. If Xk+1 ∈ Rk
    If Xk+1 is XD
        REWARD = V∈N
        DONE = True
    Else
        REWARD = -Wk
        DONE = False
    Else
        REWARD = 0
        DONE = True
6. Return: REWARD, DONE
End
    
```

In the designed networkη, the computational process for one transfer is the conditions, the environment to get respective REWARD () and DONE () subjected to network conditions of node connectivity. The process computes the transaction between node Nk→ Nk+1 as T (Nk→Nk+1). Typically, REWARD is the value returned by the algorithm that signifies encouragement if it is positive REWARD and discouragement if it is negative REWARD towards selecting the same route in the next transaction. In contrast, DONE is a final state which signifies either FAILURE or SUCCESS before restarting the next transaction. The respective values of Rk from the Nk set and Xk+1 from Rk is obtained.

Further, firstly it checks that Xk+1 is an element of Rk. If this condition is found to be true then, it checks whether Xk+1 is X_D (destination node number) or not. If it is true, then a large natural number is assigned as REWARD, and the transaction state DONE is set to True means it goes to the next iteration. Otherwise, a negative value is assigned to REWARD, and the transaction state DONE is set to False, which means the transaction further continues. In case Xk+1 does not belong to Rk, then zero is assigned as REWARD, and Transaction state DONE is set to True for the further transfer iterations of the transaction.

3) *Modeling environment:* The typical environment modeling for Net-AI Gym mimics Fig. 1, and corresponding constructs are mapped as {E, Ag}: → {O, R, Ac} and in the case of network, the action (Ac) is mapped to movement of the packets from the node (Ni) to node (Nj) as shown in Fig. 5.

Typically, there are many state or observations between starting observation to the terminating observation, and the set of $\forall Ob \in \{Ob\text{-start, Ob-next} \dots Ob\text{-Terminating}\}$ is known as one episode. In the custom design of NetAI-Gym, one episode (Eps) is a journey of a packet (P) from the source node (Ns) to the destination node (Nd), and the Eps ends when either the packet drops or it reaches the Nd. The underlying architecture of the environment adopts the Markov decision process and works in a stochastic manner following the finite state machine, as shown in Fig. 6.

In Fig. 5, the states set {S1, S2, S3 ...Sn ... Sd} is mapped with the respective nodes set {N1, N2, N3 ...Nn ... Nd}, and another state considered is Done. The possible transitions are: {Transfer, Drop, Reset, Delivered}. The network behavior exhibits randomness as the weight of the links varies due to various noises, whereas the route establishment process is entirely in control of the agent. Therefore, the Markov decision process (MDP) is justified as MDP is suitable for scenarios where the possible outcomes are influenced jointly by random variables and decision-makers. The purpose of this particular MDP is also to find the best policy (π) for the decision-maker, such that $Nk = \pi(N_{k-1})$ where the transfer gives the maximum reward.

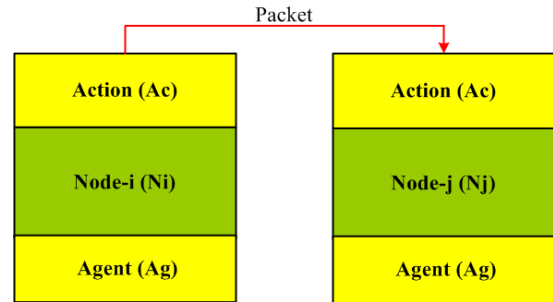


Fig. 5. Movement of Packets: Action from Ni to node Nj.

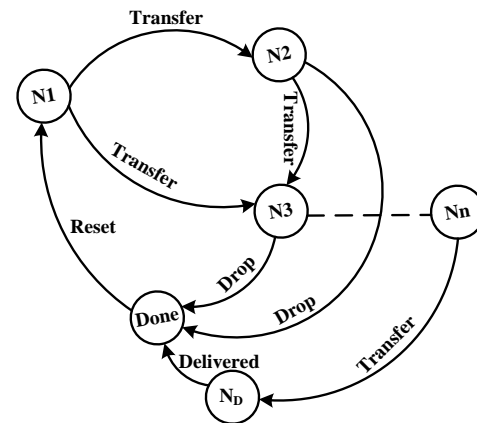


Fig. 6. Finite State Machine for Environment-Net-AI Gym.

However, since randomness is involved, a maximum REWARD cannot be ensured for every transfer. Hence, the model is designed to REWARD a huge positive value when the destination state is reached. As shown in Fig. 5, the packet may drop from any node. Hence a transition is defined from all states to DONE state as a DROP. This transition represents that any node may DROP the packet. The delivered transition is defined only from the destination state to the DONE state. This transition represents that the transaction is successful only when the packet reaches the destination node. Once it is DONE, the network must handle the next packet. The RESET transition from the DONE state shows this to the first state (Ns).

Generally, the best policy π is a function or a transformation that outputs the next state when the present state is given. This is done by either a lookup table or a function approximator to save space. The agent has to find π . This environment is being modeled to help the agent to find π efficiently. For that to happen, the model must be designed efficiently so that maximum reward is awarded when the agent reaches the destination, and it should discourage alternative non-efficient policies. Hence, the model is designed to award negative rewards when the agent tries to follow the longer path. The agent keeps exploring until no better policy exists compared to the present policy. And due to this, a well-programmed agent always finds the best policy. The environment is written so that there will always be a better policy with higher reward as long as the agent finds the best policy. The environment also ends when the agent makes a mistake. Thereby allowing the agent to learn how to ensure the packet is not dropped. All these are written keeping in mind both ML-based as well as rule-based agents. The environment is modeled so that performance doesn't deteriorate even if we scale the model. The model is made highly scalable since the overall state machine architecture is quite simple. The randomness is also modeled to simulate real-world scenarios of network disturbances. Overall, this model simulates a real-world computer network as realistically as possible.

IV. IMPLEMENTATION OF NET-AI GYM ON OPEN-AI GYM

This section presents a detailed discussion on the modeling and implementation design of the proposed RL environment for networking, namely 'Net-AI GYM'. The discussion first highlights the computational ecosystem required an intrinsic function for Net-AI Gym development followed by algorithmic steps and discussion.

A. Ecosystem for the Implementation

A professional virtual environment management tool, Anaconda is used to build the custom version of the Open-AI Gym library as Net-AI Gym. Anaconda helps to organize the required libraries, including i) well) core Python-3.8 for scripting, ii) Pandas to acquire and handle data, iii) NumPy for handling complex matrix manipulations, iv) Matplotlib for visualization through plotting. Apart from these packages, the essential package used is a NetworkX for building and managing network representation using graph theory. Since the Net-AI Gym is aimed to be used for solving various network-related problems using reinforcement learning. Therefore, to evaluate the performance of the designed agents, the stack of computational ecosystem preferred is as in Fig. 7.

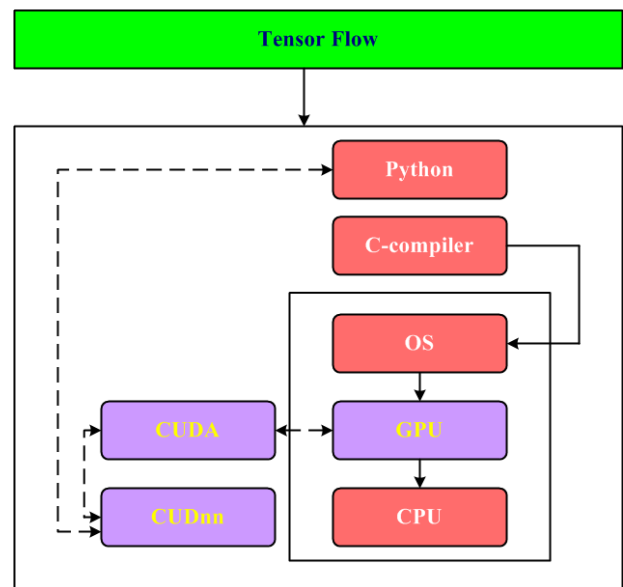


Fig. 7. Stack of Computational System for ML.

The above stack requires intrinsic operational support for complex and large matrix manipulations. Therefore, to speedup the training process, Nvidia-GTX architecture supported GPU issued along with a CPU with the Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz with 16 GB of RAM, a best-suited trailing architecture to the next best possible in the market. TensorFlow is the neural network library by Google, used to create agents to work with the environment. This can even be used to benchmark the environment; hence we need it at the top layer. As shown in Fig. 7, there are some optional components, and there are some compulsory components. A GPU is used to handle complicated matrix operations and makes the program run faster as most of the ML operations are mainly Matrix multiplication. However, even the CPU can handle matrix operations, and GPU is not a must. However, if A GPU is being used, TensorFlow must access it. CUDA is a c library given by Nvidia, allowing the C compiler access to GPU. To allow python to access GPU, we must install the CUDnn package, which acts as a wrapper to CUDA for python. If GPU access is allowed in python, CUDA will automatically access GPU.

The implementation process can proceed once the above environment is set up. To implement this environment, the NetworkX library in python is mandatory. This is required since the proposed study uses approach graph theory in this implementation. NetworkX provides an excellent source of graph theory implementation and calculations. The Graph (G) in the NetworkX contains many nodes, and each node can be treated as any bashable object. In our case, A node is nothing but an integer representing the node number. Each connection can have weight. Even though bidirectional weight is allowed in NetworkX, this feature is not being used in this implementation. During implementation, the weight from Node A to B and vice versa should be the same. This operation is ensured programmatically. NetworkX is the best-suited library to implement this environment, and more information can be stored in each node and edges by using various bashable objects in future work.

B. Intrinsic Function Development

The customization of the proposed Net-AI Gym environment includes the typical process as in Fig. 8. The flow of the environment is as shown in figure x. The init function is executed only once since the environment is initiated only once. The reset function is executed at the end of every episode. Each episode represents a transaction, i.e., the movement of a packet from the source node to the destination node. The step function is executed in a loop till the transaction is over. The step function represents a transfer, i.e., the movement of a packet from one node to another. The render function is optional as it is used only to visualize the network and transaction. Only when the learning needs to be monitored, the render function is activated.

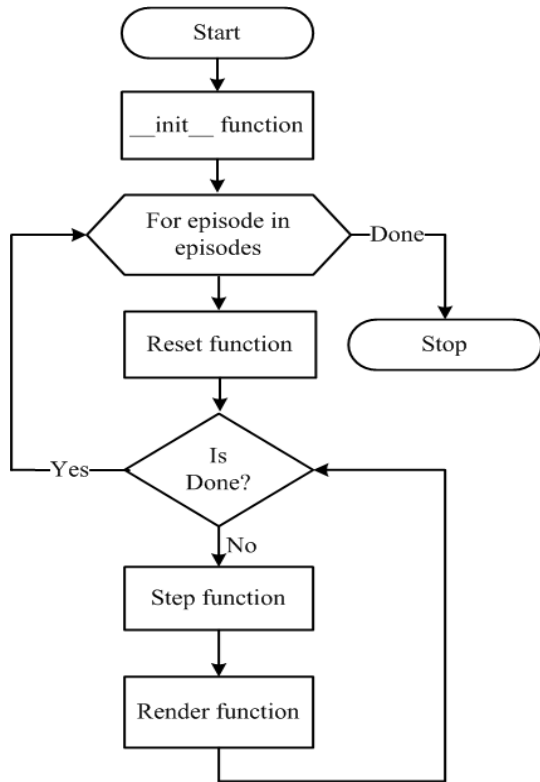


Fig. 8. Flow of Environment for Net-AIGym.

4) *Init function*: Init function is executed at the beginning of the environment to initiate it. The following algorithm represents the working of the init function.

<p>Algorithm-2: Initiation of Environment</p> <p>Input: η Output: Network N Start</p> <ol style="list-style-type: none"> 1. Initialize N 2. Length = len(η) 3. $N = \{N1, N2, N3...Nn\} = \text{createNodes}(\text{Length})$ 4. $\text{initalWeights} = \text{randomWeights}()$ 5. $\text{map}(N \rightarrow \eta)$ 6. return N <p>End</p>
--

As shown in algorithm2, environment initiation happens by copying the nodes from the data structure to the network graph. Every vector in η is mapped to a node in the network graph. This step is done to ease network operations as the network has built-in functions for network operations. As mentioned earlier, the init function is executed only once. Even if the network gets reset, the network structure stays the same, and nodes stay the same. So only those variables are initialized in this function.

5) *RESET function*: The RESET function is executed on completion of every episode. Meaning, once the packet either drops or reaches the destination, the reset function is executed. During the RESET, all those variables are changing during the execution of the environment.

<p>Algorithm-3: Environment RESET</p> <p>Input: None Output: Network N Start</p> <ol style="list-style-type: none"> 1. $Nc = Ns$ 2. Reward = 0 3. Weights = initalWeights 4. done = false 5. return N <p>End</p>

All the variables are reset back to the initial state in the RESET function as shown in the algorithm. DONE is set to false as, during the beginning of the environment, it is not done. The weights are initiated again. The network starts with the initial node itself; the current node (Nc) is set to the source node (Ns) as in the beginning. Since the DONE function is set to false in the RESET function, it can be analyzed from the flowchart; both the STEP and RENDER functions are executed at least once.

6) *STEP function*: The STEP function is where the actual transaction happens, as shown in algorithm 1. After every step next state, the REWARD and DONE are returned to the agent.

<p>Algorithm-4: STEP function</p> <p>Input: An (Action) Output: Reward,Done,next state Start</p> <ol style="list-style-type: none"> 1. $Nk = Nc$ 2. $Nk+1 = \text{action}$ 3. Reward,Done = $T(Nk, Nk+1)$ (ref ALG1) 4. Next state = $Nk+1$ 5. return Reward,Done,next state <p>End</p>

The REWARD needs to be given after every step, be it positive or negative. The positive REWARD encourages the agent to follow a similar policy, whereas the negative reward discourages the agent.

7) *RENDER function*: This function is used to visualize the output of the environment. Every step taken can be visualized. However, if the aim is to simplify the output analysis, this function can be disabled to save time.

Algorithm-5: RENDER function

```
Input: Graph (G), Current Node (Nc)
Output: None
Start
1. for every Nk ∈ G
2.   if Nc == Nk
3.     Nk.color = red
4.   else
5.     Nk.color = blue
6. plot(G)
End
```

The render function displays the graph on the screen. If the environment is being used in a Jupyter notebook, the plot needs to be cleared manually.

V. RESULT ANALYSIS FOR VALIDATING NET-AI GYM

The default Net-AI Gym with five nodes rendering of the environment is shown in Fig. 9.

Similarly, Fig. 10(a) and Fig. 10(b) illustrate the rendering of the Net-AI Gym with 50 and 100 nodes, respectively, to show the flexibility and scalability of the Net-AI Gym.

Fig. 11 shows the benchmarking of the Net-AI Gym environment with the stable-baselines benchmarking tool.

As shown in Fig. 11, the environment performs well, and an agent can find the path in it. Once the implementation is completed, a stable baselines library [25] is used to benchmark this environment. The agent can get maximum reward only when it finds the best path. For simple 6 node environments, the agents find the best path in 15 episodes. For a complex environment with 100 nodes, the agent finds the best path in 500 episodes.

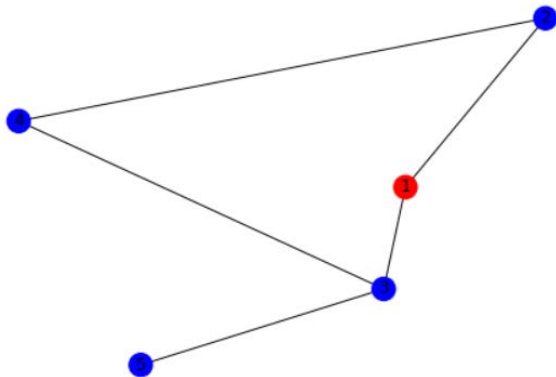


Fig. 9. Rendering of the Environment.

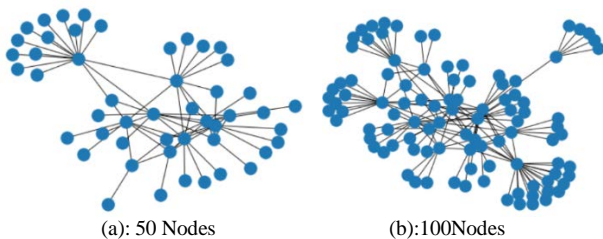


Fig. 10. (a), (b) Rendering of the Environment with 50 and 100 Nodes.

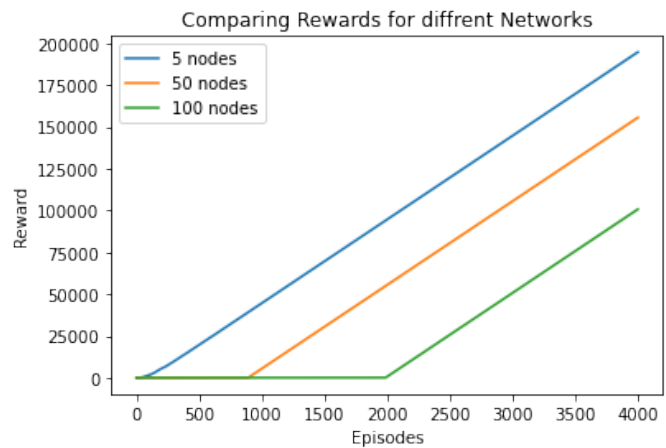


Fig. 11. Baseline Benchmark of Reward Vs. Episodes.

VI. CONCLUSION

The Open-AI Gym is a de-facto toolbox that provides numerous ready environments to test the agent algorithm's validity. A thorough investigation reveals that the appropriate environment for solving network-related problems is not available to date either by the Open-AI Gym core team or by a third-party contribution. Due to a lack of environment, the potential and advantages of RL-based agents can be fully utilized in networking problems. Therefore, the proposed study has proposed a novel approach of a customized networking environment to support RL-agent to be implemented and tested to solve various networking problems. Our future research problem is to design and develop an optimal routing algorithm for the generic network using reinforcement learning that demands a suitable environment to check the validity of the designed agent. Therefore, the need for an effective and scalable environment, Net-AI Gym, arises. The process of designing Net-AI Gym involves a setting-up stack of computational systems for ML and building a customized function. This function includes `__init__`, `Reset`, `Step`, `Render` in the core reposit of Open-AI Gym by adding procedures, such as `Transfer`, `Transaction`, `Delivered`, `Dropped`, which are as per the requirement of the network routing. The production stage includes registration of environment, re-building Open-AI Gym with registered Net-AI Gym. Finally, the Net-AI Gym environment validation is performed for scalability and proper functioning with default 5 nodes, 50 and 100 nodes. The synchronized support of NetworkX in Net-AI Gym renders the network's visualization successfully and benchmarked with different numbers of nodes 5, 50, and 100 for reward Vs. Episodes analysis shows a stable pattern. Thus, the design and construction of Net-AI Gym provide a suitable platform to evaluate network routing agent algorithms.

REFERENCES

- [1] A.H. Klopff, "Drive-reinforcement learning and hierarchical networks of control systems as models of nervous system function", *International Journal of Psychophysiology*, Vol. 1(25), pp. 42-3, 1997.
- [2] R.S. Sutton, A.G. Barto, "Reinforcement learning: An introduction", *MIT press*, 2018.
- [3] C. Sun, H. Duan, "Markov decision evolutionary game theoretic learning for cooperative sensing of unmanned aerial vehicles", *Sci. China Technol.* Vol. 58, pp. 1392-1400, 2015.

- [4] M. Hüttenrauch, S. Adrian and G. Neumann, "Deep reinforcement learning for swarm systems", *Journal of Machine Learning Research*, Vol. 20(54), pp.1-31, 2018.
- [5] R. Ghoul, J. He, S. Djaidja, M. A.A Al-qaness, and S. Kim, "PDTR: Probabilistic and Deterministic Tree-based Routing for Wireless Sensor Networks", *Sensors*, 20(6), pp.1697, 2020.
- [6] J.H. Drake, A. Kheiri, E. Özcan, and E.K. Burke, "Recent advances in selection hyper-heuristics", *European Journal of Operational Research*, Vol. 285(2), pp.405-428, 2020.
- [7] Z. A.Aghbari, A.M. Khedr, W. Osamy, I. Arif and D.P. Agrawal, "Routing in wireless sensor networks using optimization techniques: A survey", *Wireless Personal Communications*, pp.1-28, 2019.
- [8] H. Wang, N. Liu, Y. Zhang, "Deep reinforcement learning: a survey", *Front Inform Technol Electron Eng*, Vol. 21, pp. 1726-1744, 2020.
- [9] J. D. Ye and M. Zhang, "A Self-Adaptive Sleep/Wake-Up Scheduling Approach for Wireless Sensor Networks," in *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 979-992, March 2018, doi: 10.1109/TCYB.2017.2669996.
- [10] G. Künzel, L. S. Indrusiak and C. E. Pereira, "Latency and Lifetime Enhancements in Industrial Wireless Sensor Networks: A Q-Learning Approach for Graph Routing," in *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5617-5625, Aug. 2020, doi: 10.1109/TII.2019.2941771.
- [11] R. Ding, Y. Xu, F. Gao, X. Shen and W. Wu, "Deep Reinforcement Learning for Router Selection in Network With Heavy Traffic," in *IEEE Access*, vol. 7, pp. 37109-37120, 2019, doi: 10.1109/ACCESS.2019.2904539.
- [12] F. Li, X. Song, H. Chen, X. Li and Y. Wang, "Hierarchical Routing for Vehicular Ad Hoc Networks via Reinforcement Learning," in *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1852-1865, Feb. 2019, doi: 10.1109/TVT.2018.2887282.
- [13] J. Wu, M. Fang, H. Li and X. Li, "RSU-Assisted Traffic-Aware Routing Based on Reinforcement Learning for Urban Vanets," in *IEEE Access*, vol. 8, pp. 5733-5748, 2020, doi: 10.1109/ACCESS.2020.2963850.
- [14] Z. Jin, Q. Zhao, and Y. Su, "RCAR: A Reinforcement-Learning-Based Routing Protocol for Congestion-Avoided Underwater Acoustic Sensor Networks," *IEEE Sensors Journal*, vol. 19, pp. 10881-10891, Nov. 2019.
- [15] V. Di Valerio, F. Lo Presti, C. Petrioli, L. Picari, D. Spaccini and S. Basagni, "CARMA: Channel-Aware Reinforcement Learning-Based Multi-Path Adaptive Routing for Underwater Wireless Sensor Networks," in *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2634-2647, Nov. 2019, doi: 10.1109/JSAC.2019.2933968.
- [16] A. A. Bhorkar, M. Naghshvar, T. Javidi and B. D. Rao, "Adaptive Opportunistic Routing for Wireless Ad Hoc Networks," in *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 243-256, Feb. 2012, doi: 10.1109/TNET.2011.2159844.
- [17] J. Dowling, E. Curran, R. Cunningham and V. Cahill, "Using feedback in collaborative reinforcement learning to adaptively optimize MANET routing," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 3, pp. 360-372, May 2005, doi: 10.1109/TSMCA.2005.846390.
- [18] C. Yu, J. Lan, Z. Guo and Y. Hu, "DROM: Optimizing the Routing in Software-Defined Networks With Deep Reinforcement Learning," in *IEEE Access*, vol. 6, pp. 64533-64539, 2018, doi: 10.1109/ACCESS.2018.2877686.
- [19] Xu, C., Zhuang, W. and Zhang, H., 2020, October. A Deep-reinforcement Learning Approach for SDN Routing Optimization. In *Proceedings of the 4th International Conference on Computer Science and Application Engineering* (pp. 1-5).
- [20] He, X., Jiang, H., Song, Y., He, C. and Xiao, H., 2019. Routing selection with reinforcement learning for energy harvesting multi-hop CRN. *IEEE Access*, 7, pp.54435-54448.
- [21] H. Zhang, D. Zhan, C. J. Zhang, K. Wu, Y. Liu, and S. Luo, "Deep Reinforcement Learning-Based Access Control for Buffer-Aided Relaying Systems With Energy Harvesting," *IEEE Access*, vol. 8, pp. 145006-145017, Aug. 2020.
- [22] Y. Liu, D. Lu, G. Zhang, J. Tian and W. Xu, "Q-Learning Based Content Placement Method for Dynamic Cloud Content Delivery Networks," in *IEEE Access*, vol. 7, pp. 66384-66394, 2019, doi: 10.1109/ACCESS.2019.2917564.
- [23] S. Wang and Y. Shin, "Efficient Routing Protocol Based on Reinforcement Learning for Magnetic Induction Underwater Sensor Networks," in *IEEE Access*, vol. 7, pp. 82027-82037, 2019, doi: 10.1109/ACCESS.2019.2923425.
- [24] W. Jin, R. Gu and Y. Ji, "Reward Function Learning for Q-learning-Based Geographic Routing Protocol," in *IEEE Communications Letters*, vol. 23, no. 7, pp. 1236-1239, July 2019, doi: 10.1109/LCOMM.2019.2913360.
- [25] A. Hill and A. Raffin, M. Ernestus, and A. Gleave, A. Kanervisto, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, Y. Wu, "Stable Baselines" in *GitHub repository on GitHub*, 2018, (online: <https://github.com/hill-a/stable-baselines>)

Service Outages Prediction through Logs and Tickets Analysis

Sunita A Yadwad¹
Research Scholar
Department of CS and SE
Andhra University
AP, India

Dr V. Valli Kumari²
Professor
Department of CS and SE
Andhra University AP
India

Dr S Venkata Lakshmi³
Assistant Professor
Department of CSE
GITAM (Deemed to be University)
AP, India

Abstract—Service outage or downtime is a growing challenge to the service providers and end users. The major cause for the unavailability firstly is failure of equipments and applications at various places and secondly failure for proactive diagnosis and rectification. The system activities that are logged and the response of customers and providers in the form of trouble tickets could be studied for minimizing network faults. The downtime can be reduced when the failures are predicted well in time and proactively corrected. Accurate prediction of faults helps in responding to downtime even before the customer tickets are raised or network trouble is encountered. Most of the research focuses on trouble shooting through forecasting the quantity of trouble tickets using the historical ones. If these tickets can be supported with the warning in the form of Syslogs and the technical support of network tickets the predictive models would be more efficient and accurate. Dynamic and truly adaptive machine learning algorithms are essentially required for processing the torrent of data and formulating predictions based on the trends and the patterns existing in it. The work refers to i) identifying number of trouble tickets that are related to the device a few days before the network component fails, ii) predicting fault will occur in broadband networks. Lasso and Ridge regression are used for the first and Bayesian structural time series analysis and prophet are used for the latter.

Keywords—Failure prediction; linear regression technique; network fault prediction; lasso; ridge regression; Bayesian structural time series; prophet

I. INTRODUCTION

Internet is an indispensable service for mankind. Owing to its vivid and continuous usage for various purposes, life cannot be imagined in absence of this service. Detection and minimization of service unavailability or downtime is the need of the hour and a challenging task for the service providers. There are several reasons leading to the service outage.

One of the major reasons is the failure of equipments, service or application. Networks are complex due to rise in demand of novel and diversified applications. As these applications are provided by various vendors and providers it is difficult to detect network failures and diagnose their causes. The data for the detection can come from various sources like Social networking data, Syslogs, Customers tickets, Signal measurement. Customers internet usage data and the network trouble tickets. The data should have been harnessed using

several machine learning algorithms for fault prediction in networks.

The motivation for the present work comes from literature consisting of research papers that focus on fault prediction in networks using different techniques and datasets. The papers surveyed work on fault detection in two major ways:

- 1) Use quantity of customer trouble tickets with time series predictive models for prediction of the quantity of faults.
- 2) Using of System logs, Signal Measurements, data from social networking services and internet usage data that is generated by various network components for prediction of likelihood of the components being faulty in near future.

This paper proposes comparison of several machine learning techniques that help in assessing the priority and intensity of the trouble reports and choose the most optimal solution for investigation. The approach is used for detection and prediction of faults in an efficient and improved manner. To achieve the above goals for network services, the paper uses historical data obtained for the selected equipment several days before the failure which is made up of i) the customer trouble tickets used by the customers to inform about services affected, ii) Network trouble tickets which are technical information about break downs and services interruptions noted and iii) syslogs which are event logs, warning and alarms generated by the equipments.

The model proposed in the paper firstly effectively forecasts the count of customer trouble tickets for coming days using structural time series analysis overcoming all error limitation of ARIMA, GARCH, ARMA model and then uses range of warnings obtained from the historical data several days prior to the failure to formulate a pattern which can effectively predict the what happens to the equipment the following day. When a certain range of cumulative warnings are observed, that the equipment failure will happen the next day is predicted.

Due to confidentiality, the data presented in this paper are masked. The credibility of the model is supported by comprehensive tests.

Owing to confidentiality reasons data presented has been masked and model credentials are supported by several tests. The remaining portions of the paper are categorized in the

following sequence: Section 2 which immediately follows the introduction describes the related works referred and the motivation, methodology is explained in Section 3, Section 4 discusses both the regression and the forecasting results, and Section 5 is the concluding remarks.

II. RELATED WORK AND MOTIVATION

Literature has in store abundant work focusing on network fault prediction using all possible machine learning techniques. The proposed techniques try to better them in terms of accuracy of prediction of faults. Most of the existing works use the following to deal with network downtime.

- They use total number of trouble tickets created for a time series network fault forecasting model.
- They use logs generated by network component or internet usage and measuring data.

This section discusses a few such works.

A. Using Customer Trouble Tickets

Addressing customer services are highly desirable for both maintaining reputation of companies and for early prediction of outage [2]. Service outage experiences have been increasing among users of telecommunication industry [3] and root cause of the failure is inability to predict the failures proactively.

The measure of customer grievance tickets which are related to follies in network can be surmised with a model in time series. The quantity of tickets having equal intervals allows discrete representation of data. The trend can be exploited for future sequence forecasting. Customer on premise network equipment faults were focused by University of Telekom in monthly, weekly and hourly intervals using Autoregressive Integrated Moving Average (ARIMA), ARMA, GARCH, Kalman filtering and multivariate recurrent neural network model [1,2,3,4]. With sufficient amount of training data, the number of customer trouble tickets that would be generated was forecasted well ahead to let the service providers govern the allocation of their workforce. All methods were tested against the CMSE (Cumulative Mean Square Error) values for prediction of quantity of faults in broad band network. ARIMA proved to be a winner.

B. Using System Logs and Network Usage Data

System logs (Syslog) are textfiles which provide an audit trail of events. Applications send information to the syslog process which stores the message in a text file in the order that they arrive [7]. Syslogs validate those tickets which are related to failure analysis and help in prior detection of network component failures. Logistic regression has been highly used for failure prediction. Logistic regression with rule based analysis has been proposed [5] to create a credible model and directly forecast future failure of network components at least four days prior to the actual occurrence. They utilized the historical data which includes the customer trouble ticket, network trouble ticket and Syslog warnings.

Rule based analysis model for the prediction of equipment failure the very next day was constructed with the cumulative sum total of the warnings and the gradients obtained from simple linear regression and best fit of line method. Challenges

are data is available on real time basis. It is huge and needs adequate storage and processing. Also the trouble ticket and Sysco follow different format for different manufacturers.

Study in paper [6] has used the classifiers like Random Forest classifier, C5.0 decision tree algorithm for bettering the forecast of network faults. The customer tickets are combined with network signal data and internet usage data to aptly describe the customer behavior and quality of the network components.

The sliding window for analysis was chosen for seven days, the obtained data in this span from customers was augmented with the internet usage data and signal measurement of network. The C5.0 decision tree and Random Forest (RF) classifiers were used for training the data. The results proved that the RF method had higher accuracy in prediction of network fault and also could estimate the importance of variable for prediction. C5.0 could present the expression leading to prediction outcome for better understanding of the potential causes of failure by network management.

In the purview of the above findings and motivation the main objectives of the paper are:

1) Obtain the warnings from the customer tickets and represent network fault with time series. The obtained trend can help forecast the future sequence in time series. So far the authors have proved ARIMA [1-4] is the best but prove Bayesian structural time series and prophet techniques give better result than ARIMA in warning prediction.

2) With Lasso, Ridge and Elasticnet regression technique and best-fit line methods predict the equipment failure atleast 9 days in advance. The pattern in which warnings occur is studied on the daily to set a trigger well in time to take necessary actions.

The above cited objectives are achieved through experimental results.

Further the work can be extended to improve the prediction of network fault using Hidden Markov model . Data training can be done by applying powerful approach of HMM and Bayesian network model in the decision function and create the fault detection method to find whether the aggregated log data is normal or failures have been detected. HMM is able to identify the important features and list decision rules that describe the relation among selected features from data aggregation of customer tickets, network tickets and System logs. The implementation is beyond the scope of the paper and is presented in subsequent publications. of th. The HMM is trained for normal non outage records.

III. METHODOLOGY

Historical data is extracted from the grievances of customers as Customer Trouble tickets (CTT), the Network Trouble Tickets (NTT) form service providers and Syslog for a given enterprise.

CTT gives customers complaints about interruption of service whereas NTT gives technical details about breakdown of equipment.

Syslog contains event logs alarms and warnings generated by each equipment. As the network elements grow, there is huge volume of complex log data. Information in the logs needs to be extracted efficiently and precisely for trouble shooting and maintenance. Records are matched based on the equipments id. Equipments could fail due to several reasons. If the reasons are climatic like weather, power vandalism or theft it is filtered out.

By matching the tickets data from customer, the failure records of the equipments can be identified. The faults listed in customer trouble tickets are cross verified with the causes in Network ticket data which has detailed elucidation about the failure of the equipment. This is followed by analysis of the warnings generated by the equipment. Using Sequential Query Language (SQL) there is extraction of all important service failure related features causing the network equipment malfunction. The selected features from the historical data are listed below in Table I.

The block diagram in Fig. 1 represents the architectural model of the outage detection through fault forecasting and prediction.

The data collected from all the three sources is firstly stored using HDFS and later distributed across all the clusters. Each and every node processes the data using SQL in Hive. Due to presence of SQL there is no need to group or reduce the data. The native function of SQL classifies and processes the data all together to determine the total number of occurrences for the number of warning per equipment. The original raw data for very equipment in Syslog is more than million records as one month of data is used for the prediction model. The mappers key/value pairs are created with each Warning type and the respective counting values. The reducers are set to aggregate counting values giving the daily accumulative total for each Warning type and for each equipment. This in turn serves as the selected feature related to communication failure due to the failure of the equipment. These features selected are in turn fed to the regression model.

On the other hand the total numbers of customer tickets that are accumulated per day are fed to the BSTS a prophet time series model as fault data for further forecasting for coming times.

A. Regression Model

Large amount of data is collected by the equipment making the detection of failure time taking. Simple and effective methods are required for prediction with the real time and continual data. Linear regression model has been used for prediction of faults and the paper [5] has achieved accuracy of predicting faults days prior to the occurrence. The goal in this paper is to predict faults at least nine days prior to failure for proper preventive measures and continued service.

In linear regression theory before constructing the model a relationship is determined between the independent predictor variable and the dependent predicting variable. The predictor variable X is one on which prediction is based and the criterion variable Y is the one used to predict. Regression line which is a straight line is formed when Y gets plotted in terms of X as a function.

TABLE I. SELECTED FEATURES FROM HISTORICAL DATA

Customer Trouble Tickets	Network Trouble Tickets	Syslog
Date by Customer	Date by Network	Date of Syslog
Fault of Equipment E1	Causes of Fault of E1	Types of Warning for E1
Types of Equipments	Types of Equipments	Types of Equipments
Equip id	Equip id	Equip id

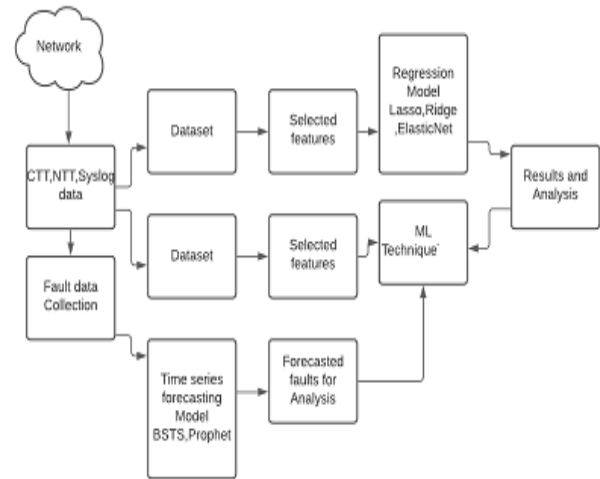


Fig. 1. Fault Forecasting and Prediction Model.

Regression formula can be written as.

$$Y = AX + C \tag{1}$$

Here A is the gradient Y and X are predicted outcome and predictor variable.

In [5], simple linear regression considers cumulative total warnings for previous days to predict what happens the next day. The variable Y is cumulative total of warnings and X represents the day the warnings will be recorded. The regression line is computed by taking values of Y against X. The value of square of correlation coefficient R^2 is computed which depicts the variance of one variable over the other. Its range varies from 0 to 1. As the value nears 1 the data has stronger relationship and determines the certainty of the predictions.

There are several cases where the classical linear regression model doesn't handle data well and accuracy can be further improvised with dimension reduction or regularization. In the Ordinary Least Squares (OLS) approach, variance and bias can be reduced with an approach called Regularization which is beneficial for the improvement in the predictive performance. The process adds information to solve the issues of ill posing of problems and prevent the over-fitting. Commonly used regularization method includes adding a constraint to the loss function

$$\text{Regularized Loss} = \text{Loss Function} + \text{Constraint}$$

Most popularly used forms of constraints in regularization are the Ridge Regression, the Least Absolute Shrinkage and

Selection Operator (Lasso Regression) and the Elastic Net regression. The R package used for implementation of regularized linear models is *glmnet*. Elastic Net can be tuned with function called *caret*. Ridge regression minimizes the sum of square residuals and penalizes the size of the estimates of the parameter towards a shrink zero. Lasso is identical to Ridge conceptually. It adds penalty to non-zero coefficients like Ridge regression penalizes the sum of the squared coefficients (L2 penalty) and the Lasso penalizes the absolute values sum (L1 penalty) [8, 9]. A new regression technique called Elasticnet is combination of both the regression methods to even out what is best in them. The combined penalties of both the regressions when tuned by cross validation leads to minimization of the loss function.

Linear regression equation looks like:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \dots + \theta_n X_n \quad (2)$$

The regularized regression methods can be understood in terms of why and how they are applied to ordinary least squares. The OLS regression mainly tries to find a hyper plane which minimizes the sum of squared errors between a predicted response and observed values [10, 11, and 12].

The cost function for ridge regression is

$$\text{minimize } (SSE + \lambda \sum_{j=1}^n \beta_j^2) \quad (3)$$

Here we come across an extra term, which is known as the penalty term. The λ in the equation here is actually represented by the alpha parameter in Ridge function [8, 9, and 10]. The penalty term here can be controlled by varying the values of alpha. When alpha takes higher values, the penalty term becomes bigger and therefore the magnitude of coefficients is smaller. So, we can see that there is a slight improvement in a model because the value of the R-Square increases.

Lasso regression method selects only few features while reducing the coefficients of other features to zero. This feature selection property does not exist in ridge.

Lasso regression is similar to ridge except that we swap L norm for L_1 norm. Instead of adding squares the absolute value is added for further improvement in the R^2 value and increase in the fitness of the mode.

$$\text{minimize } (SSE + \lambda \sum_{j=1}^n |\beta_j|) \quad (4)$$

Elastic net uses both L1 and L2 penalty term, therefore its equation look like as follows:

$$\text{minimize } (SSE + \lambda_1 \sum_{j=1}^n |\beta_j| + \lambda_2 \sum_{j=1}^n \beta_j^2) \quad (5)$$

The best model is defined as the one that minimizes the prediction error. In this case Elasticnet overpowers the other regression methods making it best fit for the prediction of faults.

B. Bayesian Structural Time Series

It is well known amongst data scientists that non-stationary series, shortage of data points and inability to distinguish trends of a time series will lead to inaccurate forecasts. Many forecast algorithms face this drawback. Structural Time Series models imitate and combine all the endemic features of regression models namely ARIMA, and the exponential smoothening

models. Bayesian Structural Time Series model is also popularly known as 'state space models' or the 'dynamic linear models' by different authors. It is a time series model that fits the overall structural change in time series dynamically. BSTS is the implementation of this model in R which is easy to use and is a function which requires a minimal mathematical understanding of the state space models [13, 14, and 15]. The benefits of Kalman filtering algorithm and Markov chain Monte Carlo (MCMC) are together used for fitting this model. Forecasts are then calculated from the predictive distribution of posterior. BSTS is more popular for its "Now casting" feature of predicting the values of time series in the present.

Prophet is a popular forecasting method which was developed at Facebook and is available as open source. It is a curve fitting approach, similar to how BSTS models the trend and the seasonality, except for the fact that it uses generalized additive models not the state-space representation for describing each component [16, 17]. For rigorous performance analysis of these methods calculating of forecast error related metrics like MAPE and RMSE is required. The findings of the study suggest that forecasting accuracy in the proposed models is better when measured against the frequently used ARIMA models.

IV. RESULTS

The fault forecasting and prediction process as described in the Fig. 2 can be divided into two major parts:

- The forecasting of faults using time series model for dealing with outage.
- The Network fault prediction nine days prior to the occurrence for proactive dealing with outage.

A. Hypothesis Testing on Lasso and Idge Regression for Failure Prediction

The dataset has approximately millionn records for months of data that has been logged about the event track records of equipment E1. The hypotheses are tested on the Equipment E1 which is known to have the maximum customer complaints. The complaints in the first month for the equipment failure as well as the consecutive two months are recorded. For consistency in the equipment failure's trend, this regression technique can be validated on other equipment of the same model .Testing is done on the equipment E1 for warning X. The choice of equipment depends on the highest customer tickets generated in the months listed. The trend developed on the equipment can be further used to test other equipments. In paper [5], where a simple regression method is used to predict equipment failure the value of R^2 is 1 for four days prior to the equipment failure E1 and R^2 was 0.6685 when checked about 9 days prior to failure of the equipment , which means smaller number of days is better in prediction of equipment's failure. To improvise the method of linear regression and to increase the accuracy of prediction the paper uses Ridge and Lasso regression methods. The value of R^2 increases significantly. The value of coefficient R^2 is 1 for four days and 0.8811, 0.8283, 0.8844 for 9 days with the Elasticnet, Lasso and Ridge regression techniques making it easier to fit for the model if the number of days is more. This helps us predict the fault nine days advance making it go easier with the service providers.

This has been tested on the equipment E1 with nine days and can be further tested on other equipments varying the number of days. The results of the regression technique is depicted in Table II.

In conclusion, the proposed method is valid for predicting failure in network equipments nine days before the actual occurrence. This means that the regression line better fits the actual data when the number of days is 9 and the methods are Lasso and Ridge regression. The cumulative warning range, gradients obtained and the other features of the dataset can be used to effectively predict the equipment failure and also estimate the variables that are important for prediction.

The value of R^2 when predicting the upcoming warnings for equipment E1 with Linear, Ridge and Lasso regression are as follows. Lasso and Ridge outperform the linear regression method in terms of best fitting of the model for prediction.

The hypothesis tested can be represented and the results can be summarized as below. The value of R^2 shows how well the regression line fits the actual data.

The paper achieved value of R^2 as 1 by linear regression and proposed their model can predict the fault 4 days prior to actual occurrence. But our regression model tested on three other methods lasso, ridge and elasticnet show the value of R^2 as 1 for 6 days and 0.8843 for 9 days. This proves that linear regression can be improvised by our model to achieve the fault occurrence warnings up to 6-9 days prior to the failure giving the service providers to take preventive actions.

Advantages of proposed Regression Model

The R^2 values that are obtained by the model are closer to 1 for not only 4 days prior to failure but show the strongest linear relationship to 6 days and 9 days prior to failure. This indicates that the regression line is fitting closely in terms of data to the actual.

B. Time Series Model for Fault Forecasting

Generally speaking, faults in the broadband network can be identified in two basic ways, first way to detect using a variety of surveillance systems that monitor network operation, another method deals with customer reports. These two sets of data are more or less overlapped. Union set includes those faults that have been reported by customers through tickets and those that have been recognized by the network supervisory system through network tickets at the same time. The prediction results of the methods mentioned in the methodology are given below. Mean Absolute Percentage Error commonly known as MAPE is used as the criterion for results evaluation. The results are represented visually in diagrams describing the relationship existing between both the original and the predicted values.

A 24 hour distribution of fault occurrence having period with 5 min interval is shown in Fig. 3. The figure shows how the faults have been distributed over a period of a day. This distribution is subjected to the ARIMA, BSTS, and Prophet forecasting time series methods for fault forecasting. This forecasting is a lead to the service providers to buck up their resources to overcome failure and avoid further outages and customer enrage.

The appearance of various faults is a statistical process with respect to time which is represented with time series, sampling of the faults data happens in every 5 min interval and the interval sequence has a cumulative feature which facilitates the series of discrete time parameter. Members of series can be forecasted. Every forecasting method is distinctive in its own way. The characteristics of every method cannot be considered a hundred percent for sure. Adequate methods are required for increase in the accuracy of the prediction.

The aim of the work is examining the prediction of faults through three different models ARIMA, BSTS and Prophet. The model is designed on the basis of monitoring of the network broadband faults analyzed in the order in which they appear. The model is described in Fig. 3.

TABLE II. RESULTS – R^2 COMPARISON

Computed value of R^2 for Equipment E1				
S.no	Regression Method	4 days	6 days	9 days
1	Linear	1	1	0.6685011
2	Elasticnet	1	1	0.8811015
3	Lasso	1	1	0.8283883
4	Ridge	1	1	0.8843964

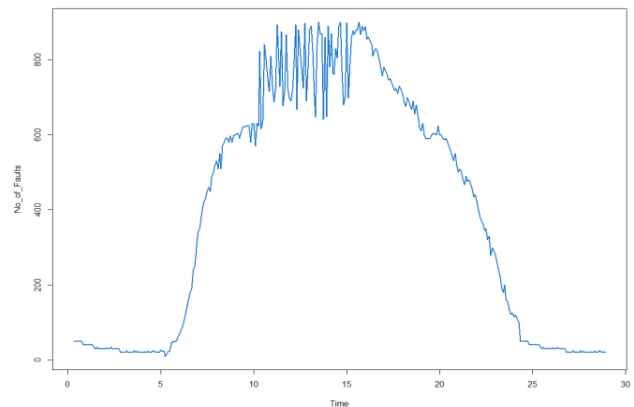


Fig. 2. Daily Fault Distribution.

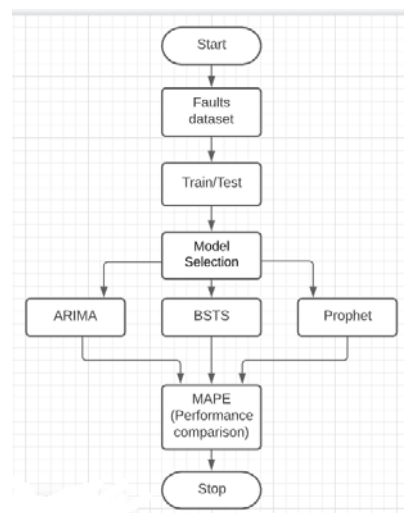


Fig. 3. Forecast Model.

Accuracy of the above mentioned models are assessed by comparing the predicted results obtained against the actual data. The results obtained in fault prediction for a randomly selected day are shown in Fig. 4 to Fig. 6 by ARIMA, BSTS and Prophet Method.

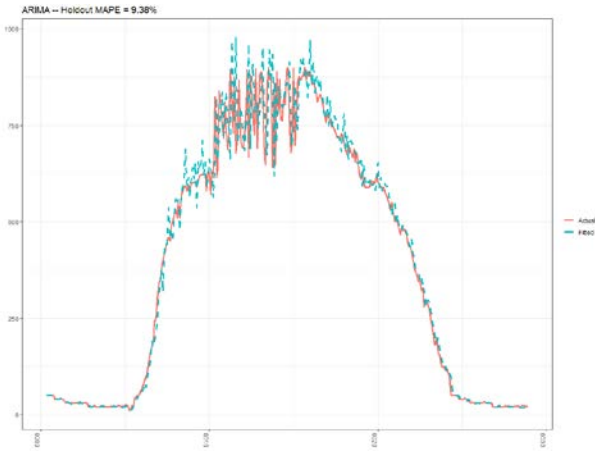


Fig. 4. Results of Forecasting using ARIMA.

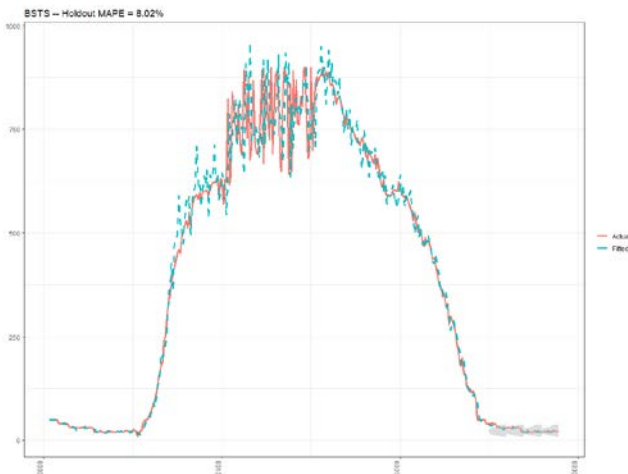


Fig. 5. Results of Forecasting using BSTS.

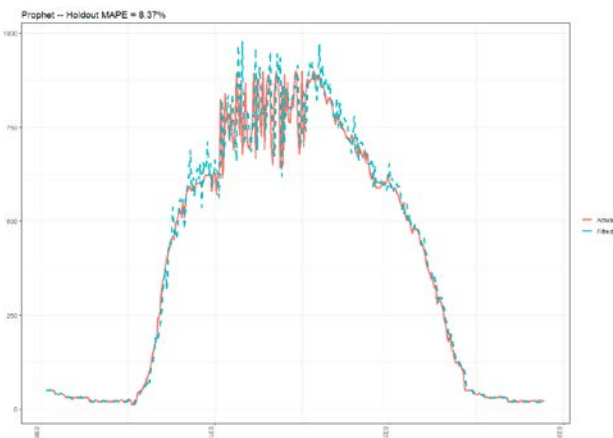


Fig. 6. Results of Forecasting using Prophet.

Focus of the paper is trying to find the best candidates for time series forecast of occurrence of faults in the broadband networks.

Comparison is made for the checking accuracy which is the difference between predicted and actual data for each of the three models. The accuracy of prediction criterion is expressed by using MAPE. The MAPE errors obtained are shown in Table III.

TABLE III. RESULTS – MAPE COMPARISON

Performance Metrics	MAPE %
ARIMA performance	9.38
BSTS performance	8.02
Prophet performance	8.37

Comparing these three models over prediction, we can conclude that better results were achieved by BSTS and prophet method over ARIMA STL. The tools show respective efficiency and achieve better results.

Advantages of Forecasting Model

The motivation for the work is that not much forecasting activities are undertaken to predict the rate of faults in a network for prevention of outages. There were some previous studies to predict fault using Kalman filters, GARCH, HMM, ARMA and ARIMA methods. Getting a formula, algorithm to with a good accuracy to predict the amount of faults is certainly a challenge. The advantage of the proposed model for forecasting is usage of Bayesian structural time series and Prophet Method which surpass the ARIMA method which has been the best contender in the surveyed papers.

V. CONCLUSION

The forecasting of faults for the broadband network accurately is the need of the hour for internet service providers. It helps them to properly strategise the future operational expenses and plan the strategies for increasing their business efficiency. The forecasted data is a means to make decisions concerned to the network maintenance, investments in terms of new equipments and proper work to resource allocation. Proactive actions can be directed in the areas which are identified to be the potential generators of network service faults. Increase quality of service to the customers is the major driving force behind the research.

Firstly the paper studies data about the equipment through customer grievances, through network tickets and the warnings of Syslog. Failures related to Equipment can be predicted to ascertain less downtime and more customer satisfaction. Algorithms namely Lasso and Ridge to process the huge amount of data through a regression technique to predict the equipment failure are proposed and implemented. They optimize linear regression and improve the model for early prediction with better accuracy. The work can be further expanded by adding additional warning types and by establishing significant relationship among the different types of equipment failure.

Secondly this paper compared three short-term prediction models to derive the number of faults that would occur in telecommunication networks namely BSTS, Prophet and ARIMA which previously has been proved to be the best method. The aim was finding the best contender for the analysis and forecasting of faults occurring. The conclusion drawn was that the BSTS and Prophet models showed higher prospects in forecasting network faults in telecommunications networks.

In future the achieved accuracy can be further improvised by expansion of the model with addition of different data sources, accumulated both from the network and other external information. These can be further assimilated with probabilistic chains of the dynamic Bayesian network and Hidden Markov Model. This can be the topic for future extension of the research.

REFERENCES

- [1] Sonny Yuhensky, Hafiddudin, Rendy Mi, et al. "Forecasting the formulation model for amount of fault of the cpe segment on broadband network PT.Telkom using ARIMA method ". In Control, Electronics, Renewable Energy and Communications (ICCEREC) , 2016 International Conference on, pages 185–191. IEEE, 2016.
- [2] Zeljko Deljac, and Gordan Krcelic, Mirko Randic , "A multivariate approach to predicting quantity of failures in broadband networks based on a recurrent neural network" . Journal of network and systems management, 24(1):189–221, 2016.
- [3] Deljac, Zeljko, Boris Spahija and Marijan Kunstic, "A comparison of traditional forecasting methods for short-term and long-term prediction of faults in the broadband networks." MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, 2011.
- [4] Zeljko Deljac and M. Kunstic. "A comparison of methods for fault prediction in the broadband networks." Software, Telecommunications and Computer Networks (SoftCOM), 2010 International Conference on. IEEE, 2010.
- [5] Lam Hai Shuan, Guo Xiaoning, Tan Yi Fei, Soo Wooi King and Lee Zhe Mein." Network equipment failure prediction with big data analytics". International Journal of Advances in Soft Computing & Its Applications, 8(3), 2016. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [6] Ji Sheng Tan, Amy Hui-Lan Lim ,Chin Kuan Ho, "Predicting network faults using random forest and C5.0," International. J. Eng. Technol., vol. 7, no. 2, pp. 93–96, 2018.
- [7] T. Kimura, K. Takeshita, M. Yokota, K. Nishimatsu, , T. Toyono and T. Mori, "Network Failure Detection and Diagnosis by Analyzing Syslog and SNS Data: Applying Big Data Analysis to Network Operations," NTT Technical Review, Nov 2013, Vol. 11, No. 11.
- [8] Rodríguez, Oldemar. (2013). A Generalization of Ridge, Lasso and Elastic Net Regression to Interval Data. 10.13140/2.1.3753.0883.
- [9] Friedman, J., Simon, N., Hastie, T, Tibshirani, R. , (2015). <http://cran.r-project.org/web/packages/glmnet/index.html> , glmnet: Lasso and elastic-net regularized generalized linear models.
- [10] Pereira, José & Basto, Mario & Ferreira-da-Silva, Amélia. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*. 39. 634-641. 10.1016/S2212-5671(16)30310-0.
- [11] Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, 1996, pp. 267–288. *JSTOR*, www.jstor.org/stable/2346178. Accessed 18 Jan. 2021.
- [12] Friedman,J, Hastie.T, Simon .N ,Tibshirani R.,(2015). glmnet,Lasso and elastic-net regularized generalized linear models. R package version 1.9-8, <http://cran.r-project.org/web/packages/glmnet/index.html>.
- [13] KIM LARSEN , "Sorry ARIMA, but I'm going Bayesian" ,April 21,2016,San Francisco ,CA.
- [14] G. Papacharalampous, H. Tyralis, D. Koutsoyiannis, "Predictability of monthly temperature and precipitation using automatic time series forecasting methods", *Acta Geophysica* 66 (4) (2018) 807–831. .
- [15] Abhinaya Ananthakrishnan "Forecasting ? Think Bayesian" .
- [16] July 8th ,2018 , End to end problem solving ,Data nerd,Facebook ,UT ,Austin.
- [17] Almarashi, Abdullah & Khan, Khushnoor. (2020). Bayesian Structural Time Series. *Nanoscience and Nanotechnology Letters*. 12. 54-61. 10.1166/nml.2020.3083.
- [18] N. Bounceur, I. Hoteit and O. Knio, "A Bayesian Structural Time Series Approach for Predicting Red Sea Temperatures," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1996-2009, 2020, doi: 10.1109/JSTARS.2020.2989218.

Recent Themes of Colombian Scientific Engineering Journals in Scopus

Marco Aguilera-Prado¹

Vice-presidency for Research
Universitaria Agustiniiana
Bogotá, Colombia

Octavio José Salcedo Parra²

Faculty of Engineering
Universidad Distrital Francisco José
de Caldas, Bogotá, Colombia

Eduardo Avendaño Fernández³

Faculty of Engineering
Universidad Pedagógica y
Tecnológica de Colombia
Tunja, Colombia

Abstract—Through a co-occurrence bibliometric and citation analysis of 1,272 texts published in the four Colombian engineering journals available in Scopus between 2014 and 2018, this paper identified that most articles belong to supply chain optimization and logistics and involve work with information that requires minimal laboratory experimentation. Works applying artificial neural networks, clustering, and genetic algorithms are also prominent. Results from researching on biomass analysis on bioenergy and sustainability are more recent and are present to a lesser extent. Most of the reference texts of the articles published come from Spanish-speaking countries and mostly cite DYNA, the European Journal of Operational Research, the Journal of Food Engineering, and Ingeniería e Investigación.

Keywords—Co-occurrence words; bibliometric analysis; bibliometrics; Colombian journals

I. INTRODUCTION

The dynamics of Colombian journals in the Scielo database show that: i) Colombian journals are the leading destination of texts with Colombia affiliation; between 2002 and 2013, 12,534 articles (81.9%) were published in these journals; ii) the journals with most articles published were Biomédica (449), Revista de Salud Pública (438) and Agronomía Colombiana (369); iii) most collaboration is between Colombian authors (56%), whereas the most prominent international collaboration was with Spain (438, 2.9%), Brazil (422, 2.8%), and the USA (418, 2.8%). Furthermore, both the titles of the journals where articles are published, and the articles' themes show that health sciences, engineering, and biology are the most significant fields of the bibliographic production of Colombian scientific research published in the region's journals [1].

Regarding engineering journals in Scopus, the searching results shows four Colombian journals under the Engineering (miscellaneous) category: DYNA (Universidad Nacional de Colombia), Revista Facultad de Ingeniería (Universidad de Antioquia), Ingeniería e Investigación (Universidad Nacional de Colombia), and Ingeniería y Universidad (Pontificia Universidad Javeriana). These journals have been publishing research and review articles in multiple areas of engineering at a national level for over 20 years and have been indexed in Scopus for 10 years.

According to SJR, a detailed look at the quartile classification shows that: i) the four journals were part of

quartile three in 2017; ii) DYNA achieved quartile two between 2012 and 2016; iii) in 2017, the classification by SJR showed that Ingeniería and Investigación (0,189) in the first place, followed by Revista Facultad de Ingeniería (0,172), DYNA (0,167), and Ingeniería y Universidad (0.161), and iv) the h-index shows DYNA (11) in the first place, followed by Revista Facultad de Ingeniería (9), Ingeniería e Investigación (7) and Ingeniería y Universidad (4), which are relatively low indexes compared to similar journals in quartile three.

It was also found that: i) generally, the Colombian academic production on engineering comes mainly from Universidad Nacional and Universidad de Antioquia, and focuses on recognized Latin American journals such as DYNA, Revista Facultad de Ingeniería, and Revista Ingeniería e Investigación [2]; ii) it is characterized for the low citation of its articles [2], and iii) about half of the texts on Colombian engineering journals in Scopus are written in Spanish, by authors from non-English speaking countries [3].

The possible reasons for why this citation indicators go along the various characteristics of scientific production may be grouped in three sets [4]: aspects of the journals (impact, language of publication, field, form of publication (conference, journal, proceeding), authors' characteristics (number of authors, authors' reputation, self-citation, national or international collaboration, affiliation, authors' sex, age and ethnicity), and specificities of the documents (text quality, novelty, popularity, and relevance of the subject matter, field characteristics, methodology, type of document, results' characteristics, use of figures, tables and appendices, metadata characteristics and references, text extension, date of publication, citation speed).

In that sense, and as part of the discussion about the characteristics of Colombian academic publications on engineering, this article identifies the most recurrent themes of the articles published in these journals as a first step to find similarities and differences with the journals ranked in the highest quartiles and thus have the way to establish particularities that explain their citation.

This article also shows which bibliographic networks are used, i.e., the source of the references (journals and magazines) to the published articles. Specifically, following academic literature for bibliometric journal reports [5], [6] the document makes a quantitative description of published articles

emphasizing on its number and citation averages and then, like in scientific networks research [7], [8], [9], [10] uses a co-words analysis to identify co-citation networks and issues from its keywords.

Following this introduction, the method and the information sources used are presented, the resources are shown, and the findings are discussed.

II. MATERIALS AND METHODS

Systematic Literature Network Analysis (SLNA), which comprises two stages, was used to analyse the articles. The first stage includes a literature review to define the universe of texts to be considered [11], [12] [13]. The study's scope and location, the search criteria, and the selection of the texts to be analysed were defined in this step. The second stage includes the analysis and the visualization of the networks. The results of this step are the maps or networks with which the proposed SLNA objective is fulfilled.

The first step showed that the working universe would include texts classified as articles published in Colombian journals of the Engineering category in the Scopus database and published between 2014 and 2018 (TABLE I). Two types of networks were built based on these articles: i) by co-occurrence of terms for keywords [14], [15], [16], and ii) by citation, to establish the sources of the references to the articles published. VOSviewer tool was used for mapping purposes [17].

TABLE I. UNIVERSE OF ARTICLES PUBLISHED PER YEAR

Journal	2014	2015	2016	2017	2018	Total
DYNA (Colombia)	189	180	151	155	44	719
Revista Facultad de Ingeniería	80	70	58	39	16	263
Ingeniería e Investigación	42	57	41	40	26	206
Ingeniería y Universidad	21	21	17	13	12	84
Total	332	328	267	247	98	1272

Source: authors based on Scopus

Word co-occurrence analysis or co-words analysis was used to build the networks, leading to finding certain similarities in a set of texts based on keywords (author or journal) of the texts to analyse, so patterns or structures for the related words may be established [15]. This, because if two words appear in a text, then there are related themes, and if two words appear in two different texts, then the texts are related.

The results below show indicators of the number of publications, citation as a way to show the scientific relevance of journals [18], [19] and keyword co-occurrence and citation mapping to establish the subject matters of recent publications in Colombian engineering academic journals.

III. RESULTS

Colombian engineering journals in Scopus show a different dynamic from the journals ranked in the same quartile: i) fewer texts have been published in the last three years than the average of Q3 journals; ii) their number of citations during the last three years is lower than the average of Q3 journals, and

iii) the citations per text for the last two years are higher than Q3 journals for Ingeniería e Investigación and Ingeniería y Universidad, and lower for DYNA and Revista Facultad de Ingeniería (TABLE II).

The comparison between the four journals selected shows differences in the number of texts published, the citations received, and their date of creation. Thus, DYNA, created in 1933, is the oldest journal and has most of the published texts. Revista de Facultad de Ingeniería, created in 1984; Ingeniería e Investigación, created in 1981, and Ingeniería y Universidad, which published its first issue in 1997, are next (TABLE III).

Furthermore, most citations are from DYNA, which has had over 200 citations for the previous three years since 2014. DYNA is followed, in order of citations, by Revista Facultad de Ingeniería, Ingeniería e Investigación, and Ingeniería y Universidad which did not exceed 100 citations in the previous three years (TABLE IV).

During the period between 2014 and October of 2018, the journals published 1,272 texts classified as articles. Most of these were published by DYNA (57%), followed by Revista Facultad de Ingeniería (21%), Ingeniería e Investigación (16%), and Ingeniería y Universidad (7%). Most of these articles (59%) have not been cited, and DYNA has the highest percentage of articles with at least one citation (TABLE V). The range of citations of the articles is between one and sixteen citations.

Thus, Colombian engineering journals in comparison with their peers in the third quartile of Scopus are below the average number of articles published and the two- and three-years average citation. The evidence shows that as of 2018, national journals have not exceeded the barrier of 540 articles published in three years, neither 200 citations in three years. This, in the medium term, if the trend is not reversed, would point to a decrease in the classification by SJR (Scimago Journal Rank) of Colombian engineering journals.

TABLE II. TEXTS AND NUMBER OF CITATIONS PER QUARTILE IN SCOPUS

Journals	Average docs. (3 years)	Average citations (3 years)	Average citations/doc (2 years)
Q1	591	2294	2.939
Q2	364	339	0.893
Q3	544	267	0.438
Q4	233	49	0.201
Average	440	767	1.158
	No. Docs. (3 years)	No. Citations (3 years)	Citations/ doc (2 years)
DYNA	532	200	0.340
Ingeniería e Investigación	149	83	0.600
Ingeniería y Universidad	62	32	0.690
Revista Facultad de Ingeniería	219	68	0.310

Source: author's calculations based on Scopus.

TABLE III. NUMBER OF TEXTS PUBLISHED IN 3 YEARS

Journal	2010	2011	2012	2013	2014	2015	2016	2017
DYNA	173	285	361	419	438	492	512	532
Revista Facultad de Ingeniería	137	253	292	276	226	213	224	219
Ingeniería e Investigación	60	126	213	202	175	132	144	149
Ingeniería y Universidad	31	46	61	69	79	73	68	62
Total	401	710	927	966	918	910	948	962

Source: author's calculations based on Scopus.

TABLE IV. NUMBER OF CITATIONS PER JOURNAL (3 YEARS)

Journal	2010	2011	2012	2013	2014	2015	2016	2017
DYNA	26	69	149	145	211	234	214	200
Revista Facultad de Ingeniería	18	52	50	44	49	60	87	68
Ingeniería e Investigación	3	13	23	31	42	58	74	83
Ingeniería y Universidad	2	5	8	7	7	15	14	32
Total	49	139	230	227	309	367	389	383

Source: Author's calculations based on Scopus.

TABLE V. PERCENTAGE OF ARTICLES WITH AT LEAST ONE CITATION

Journal	2014	2015	2016	2017	2018	Total
DYNA (Colombia)	63.5	47.8	38.4	18.7	6.8	41.2
Revista Facultad de Ingeniería	46.3	61.4	27.6	15.4	12.5	39.5
Ingeniería e Investigación	64.3	64.9	61.0	17.5	3.8	47.1
Ingeniería y Universidad	38.1	71.4	35.3	15.4	0.0	36.9
Total	57.8	55.2	39.3	17.8	6.1	41.5

Source: author's calculations based on Scopus.

The keyword co-occurrence mapping with at least five repetitions, throws 48 keywords grouped in 10 clusters: i) biodiesel, decision-making, education, logistics, methodology, optimization, supply chains; ii) artificial neural networks, grouping, costs, power consumption, ergonomics, FPGA; iii) efficiency, electric vehicles, management, mathematical models, risk assessment; iv) absorption, biomass, carbon, kinetics, pyrolysis; v) genetic algorithm, harmonic distortion, parameter estimation; vi) open code, Raman spectroscopy, hyperspectral representation; vii) CFD, finite elements, temperature, friction; viii) renewable energy, energy efficiency, distributed generation; ix) quality, simulation, dynamic systems, x) pollution, sustainability (Fig. 1).

Those clusters may be aggregated in six areas: Optimization, Humans, Energy, Materials, Simulation and algorithms, and Environment and sustainability (TABLE VI).

TABLE VI. CLUSTERS AND THEMES IN COLOMBIAN ARTICLES

N.	Cluster	Area
1	Biodiesel, decision-making, education, logistics, methodology, optimization, supply chains;	Optimization
2	Artificial neural networks, grouping, costs, power consumption, ergonomics, FPGA;	Humans
3	Efficiency, electric vehicles, management, mathematical models, risk assessment;	Energy
4	Absorption, biomass, carbon, kinetics, pyrolysis;	Materials
5	Genetic algorithm, harmonic distortion, parameter estimation;	Optimization
6	Open code, Raman spectroscopy, hyperspectral representation;	Simulation and algorithms
7	CFD, finite elements, temperature, friction;	Simulation and algorithms
8	Renewable energy, energy efficiency, distributed generation;	Energy
9	Quality, simulation, dynamic systems,	Simulation and algorithms
10	Pollution, sustainability	Environment and sustainability

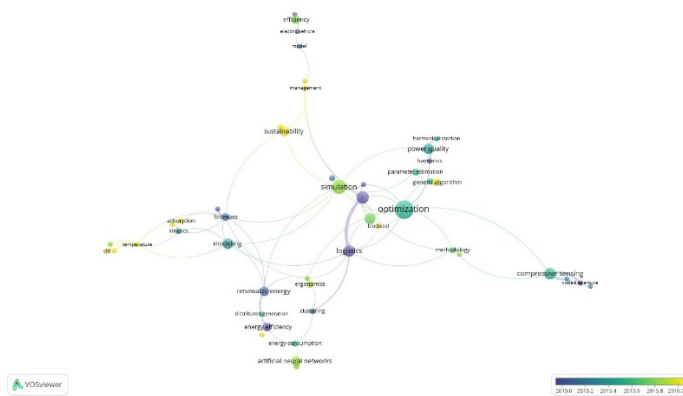


Fig. 1. Keywords Co-occurrence. Articles between 2014 and 2018.

Clusters and areas (TABLE VI) show that articles in Colombian engineering journals outline computational applications to typical engineering problems and do not deal with more recent issues such as bio or nano materials, cybernetics, geosciences or biomedical which are typical subjects in top ranked journal.

Regarding the citations by countries, the mapping shows the connection of 53 countries as sources of bibliographic references of the articles, showing Colombia, Mexico, Spain, and Brazil as the leading countries, followed by Chile, Ecuador, and Argentina, to a lesser extent (Fig. 2).

In turn, the citation per source (journal) shows that the most referenced journals by the authors of the selected articles are: DYNA (518), European Journal of Operational Research (155), Journal of Food Engineering (119), Ingeniería e Investigación (117), IEEE Transactions on Power Delivery (111), Construction and Building Materials (108) and Fuel (103) (Fig. 3).

The sources of these references (except for Ingeniería e Investigación) are associated with journals in the higher quartile, an h-indexes above 100, and articles published since the 70s. The areas covered by these journals include information systems, operations research, mathematical modelling and simulation, food science, electrical, electronic, civil, chemical, materials, and fuels engineering (TABLE VII).

TABLE VII. MOST SIGNIFICANT SOURCES OF REFERENCES OF COLOMBIAN ENGINEERING JOURNALS IN SCOPUS

Journal	Year indexed in Scopus	Journal areas	H-Index	SJR Quartile
<i>Fuel</i>	1970	Chemical Engineering Chemical Engineering (miscellaneous) Chemistry Organic Chemistry Energy Energy Engineering and Power Technology Fuel Technology	165	Q1
<i>European Journal of Operational Research</i>	1977	Decision Sciences Information Systems and Management Management Science and Operations Research Mathematics Modelling and Simulation	211	Q1
<i>Journal of Food Engineering</i>	1982	Agricultural and Biological Sciences Food Science	142	Q1
<i>IEEE Transactions on Power Delivery</i>	1985	Energy Energy Engineering and Power Technology Engineering Electrical and Electronic Engineering	152	Q1
<i>Construction and Building Materials</i>	1987	Engineering Building and Construction Civil and Structural Engineering Materials Science Materials Science (miscellaneous)	109	Q1
<i>Ingeniería e Investigación</i>	2009	Engineering Engineering (miscellaneous)	7	Q3

Source: authors

Co-words analysis shows close relationships between Colombian engineering journals and those from Spanish-speaking countries as reference sources, which is aligned with the origins and circulation of the published articles. On the one hand, most of the articles published in Colombian journals are in Spanish, which limits the access of English-speaking communities. On the other hand, if the references come from Spanish articles written in Latin America, it would be expected that the research that give rise to the articles published were about local problems and not about global coverage or those dealt with in the Scopus' Q1 or Q2 journals that have the greatest impact.

IV. DISCUSSION AND CONCLUSIONS

The findings for Colombian engineering journals in Scopus show a similar behaviour to those of Colombian journals in Scielo [1] regarding the source of the texts and the collaboration between authors. In engineering and the aggregate of Colombian journals, most texts are signed by authors affiliated to Colombia and with some participation by Spain, Brazil, and other Latin American countries. This may be a sign of preference for certain local themes and of limited

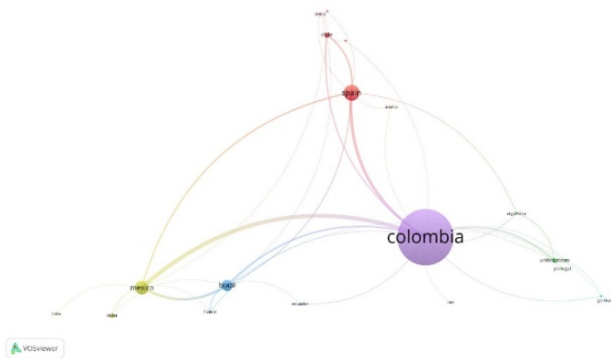


Fig. 2. Sources of the References used by Country. Articles between 2014 and 2018.

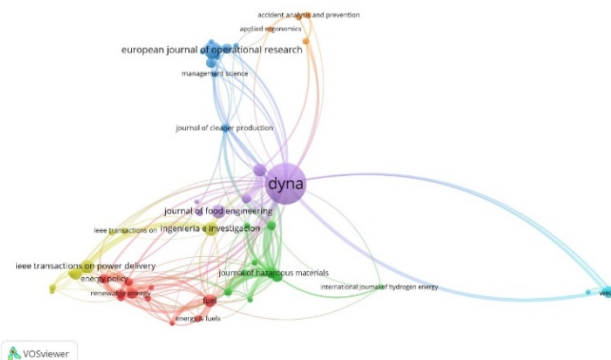


Fig. 3. Sources of References used by Journal. Articles between 2014 and 2018.

possibilities for progress on themes with greater global coverage or with particular needs (laboratories, equipment, reagents) for research.

The journal DYNA of Universidad Nacional de Colombia shows the best results on the number of articles published, the number of citations of its articles, and the number of times it is referenced in other journal articles. This may be because i) it is a traditional publication created by the first school of mines in Colombia, dating to the 19th century and which would become the Faculty of Mines of Universidad Nacional de Colombia in 1944; ii) it is the oldest publication, and iii) it is the Colombian engineering magazine indexed in Scopus with the most published articles.

Over the last few years, Colombian engineering academic journals published mostly articles on optimizing supply chains and logistics (77%), implying work with information that requires minimal laboratory experimentation. In this same vein of research that does not necessarily involve large deployments of laboratory equipment, works with applications of artificial neural networks, clustering, and genetic algorithms stand out. That may be a consequence of low investments in technology (Colombia's investment in research and development in 2018 was 0,28 % of GDP), deficiencies in higher education and its training objectives, and a low appropriation of science.

To a lesser extent (10%), texts showing the results of biomass analysis and on energy and sustainability were also published. The articles published mostly reference texts from Spanish-speaking countries. Colombia, Mexico, Spain, Chile, Argentina, Ecuador, and from the journals DYNA, European Journal of Operational Research, Journal of Food Engineering and Ingeniería e Investigación.

Thus, relevant themes of the Colombian engineering journals available in Scopus aim cover some of the conventional aspects of industrial engineering with hints of chemical engineering, mining engineering, food engineering, and artificial intelligence applications (neural networks, genetic algorithms), which contrasts with more recent engineering issues in international literature and which are present in the journals ranked in the higher quartiles (1 and 2), with higher citations: materials, physics, and astronomy, biophysics, biochemistry, ceramics, biotechnology or biomedical engineering, in which countries such as Brazil, Russia, India, and China have had significant progress [20]. This difference in subject matters may explain the low citation of articles—16 citations in the articles with the most citations in the last five years—and citations below the average for journals ranked in quartile 3.

Indeed, a research panorama emerges to clarify the reasons for those themes in Colombian journals, drawing the following hypothesis: i) the technological characteristics of the Colombian industry that includes low productivity, little national set of companies and major exports in commodities imply that the national academic literature on engineering does not deal with recent themes addressed by the international academy; ii) Colombian authors dealing with more global themes belong to international researchers networks that submit their work to international journals ranked in higher quartiles; or iii) the Colombian academy does not have the

physical capital (laboratories, instruments, servers) to conduct research that results in publications aligned with the most recently cited themes.

It is also pertinent to investigate whether there are other causes—outside research—for the articles on more recent themes not to reach Colombian engineering journals, such as the low number of volumes published, speed of response to reviews, SJR results, and the language.

REFERENCES

- [1] Maz-Machado, N. N. Jiménez-Fanjul y M. E. Villarraga, «La producción científica colombiana en SciELO: un análisis bibliométrico.» *Revista Interamericana de Bibliotecología*, vol. 39, n° 2, pp. 111-119, 2016.
- [2] J. I. Rojas-Sola y C. De San-Antonio-Gómez, «Análisis bibliométrico de las publicaciones científicas colombiana en la categoría engineering, multidisciplinaria de la base de datos Web of Science (1997-2009),» *Dyna*. 77 (164), pp. 9-17, 2010.
- [3] M. Aguilera-Prado, C. Aguirre y O. Salcedo, «Approach to Citation Determinants of Articles from Colombian Engineering Journals in Scopus,» *Contemporary Engineering Sciences*, vol. 10, n° 26, pp. 1279-1286, 2017.
- [4] I. Tahamant, A. Safipour Afshar y K. Ahmndzadeh, «Factors affecting number of citations: a comprehensive review of the literature,» *Scientometrics* 107, pp. 1195-1225, 2016.
- [5] R. P. Leone, L. M. Robinson, J. Bragge y O. Somervuori, «A citation and profiling analysis of pricing research from 1980 to 2010,» *Journal of Business Research* 65, p. 1010-1024, 2012.
- [6] C. Lokker, A. McKibbin, R. J. McKinlay, N. L. Wilczynski y B. Haynes, «Prediction of Citation Counts for Clinical Articles at Two Years Using Data Available within Three Weeks of Publication: Retrospective Cohort Study,» *British Medical Journal*, vol. 336, n° 7645, pp. 655-657, 2008.
- [7] J. Xi, S. Kraus, M. Filser y F. Kellermans, «Mapping the field of family business research: past trends and future directions,» *International Entrepreneurship Management Journal*, vol. 11, pp. 113-132, 2015.
- [8] J.-I. Hung, «Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics,» *British Journal of Educational Technology*, pp. 5-16, 2012.
- [9] R. N. Kostoff, D. R. Toothman, H. J. Eberhart y J. A. Humenik, «Text mining using database tomography and bibliometrics: A review,» *Technological Forecasting & Social Change* 68, pp. 223-253, 2001.
- [10] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma y K. Börner, «Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches,» *PLOS One*, vol. 6, n° 3, pp. 1-11, 2011.
- [11] C. Colicchia y F. Strozzi, «Supply chain risk management: a new methodology for a systematic literature review,» *Supply Chain Management: An International Journal*, vol. 17, n° 4, pp. 403-418, 2012.
- [12] F. Strozzi, C. Colicchia, A. Creazza y C. Noè, «Literature review on the 'Smart Factory' concept using bibliometric tools,» *International Journal of Production Research*, vol. 55, n° 22, pp. 6572-6591, 2017.
- [13] F. Kithous, F. Strozzi, A. Urbinati y F. Alberti, «A Systematic Literature Network Analysis of Existing Themes and Emerging Research Trends in Circular Economy,» *Sustainability*, vol. 12, n° 4, pp. 1633-1657, 2020.
- [14] J. Charum, «Generación de un sistema de información y construcción de indicadores de las acumulaciones y de las dinámicas sociales y científicas de las Red Caldas,» de *Hacer ciencia en un mundo globalizado. La diáspora científica colombiana en perspectiva*, J. Charum y J. Meyer, Edits., Bogotá, Tercer Mundo, 1998, pp. 5-40.
- [15] N. J. van Eck y L. Waltman, «How to normalize co-occurrence data? An analysis of some well-known similarity measures,» *Journal of the American Society for Information Science and Technology* 60, pp. 1635-1651, 2009.
- [16] D. H. Rodríguez y C. E. Pardo, «Programación en R del método de las palabras asociadas,» *Universidad Nacional de Colombia, Bogotá*, 2007.

- [17] N. J. van Eck y L. Waltman, «Software survey: VOSviewer, a computer program for bibliometric mapping,» *Scientometrics*, n° 84, pp. 523-538, 2010.
- [18] J. González , M. Moya y M. A. Mateos, «Indicadores bibliométricos: características y limitaciones en el análisis de la actividad científica,» *Anales Españoles de Pediatría*, vol. 47, n° 3, pp. 235-244, 1997.
- [19] D. Aksnes, L. Langfeldt y P. Wouters, «Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories,» *SAGE Open*, pp. 1-17, 2019.
- [20] B. Elango, «A bibliometric analysis of literature on engineering research among BRIC countries,» *Collection and Curation*, vol. 38, n° 1, pp. 9-14, 2019.

Book Recommendation for Library Automation Use in School Libraries by Multi Features of Support Vector Machine

Kitti Puritat¹, Phichete Julrode², Pakinee Ariya³, Sumalee Sangamuang⁴, Kannikar Intawong⁵

Department of Library and Information Science, Chiang Mai University, Chiang Mai, Thailand^{1,2}

College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand^{3,4}

Faculty of Public Health, Chiang Mai University, Chiang Mai, Thailand⁵

Abstract—This paper proposed the algorithms of book recommendation for the open source of library automation by using machine learning method of support vector machine. The algorithms consist of using multiple features (1) similarity measures for book title (2) The DDC for systematic arrangement combination of Association Rule Mining (3) similarity measures for bibliographic information of book. To evaluate, we used both qualitative and quantitative data. For qualitative, sixty four students of Banpasao Chiang Mai school reported the satisfaction questionnaire and interview. For Quantitative, we used web monitoring and precision measures to effectively use the system. The results show that books recommended by our algorithms can suggest books to students “Very interested” and “interested” by 14.5% and 22.5% and improve usage of the OPAC system's highest average of 52 per day. Therefore, these systems suitable for library automation of Thai language and small library with not much book resource.

Keywords—Library automation; book recommendation system; library integrated system; title similarity; support vector machine; open source

I. INTRODUCTION

In Thailand, libraries are the main source of knowledge in education institutions which provide resources such as books, journals, research papers, interactive media, etc. in order to support learning for people and students. In addition of Thailand, There are five types of library such as school Library, college and University, public Library, special Library and National Library which are required many librarian to management the resource such as adding new member, book data, catalogue book, giving support to clients, organizing all relevant information about books, etc. Thus, In order to support librarian, Software of library automation or Integrated library system [2][3] are the concepts of using information and communications technologies (ICT) which are designed to support and management all of manual processing task in library in order to reduce the duty of librarians to manage library work to maximum effective.

In recent years, Artificial intelligence is usually becoming a part of everyday life that impacts the way of knowledge of technology and the world. In addition, for the e-commerce sector, recommendations technique has been used widely in information agents that attempt to predict and suggest which items from data collection to user who may be interested in and

recommend favourite thing to them. The idea of technic is detecting the information process obtained by user's interaction or need according to collaborative behavior[1], the algorithms of recommendations technique represent as desirable items filtering by system from the user's past behavior which usually record and compute by purchased or selected, items previously, ratings given to those items.

However, one of the most important tasks of a librarian is to recommend interesting books for users [5] because users are not familiar with the library and lack of knowledge in accessing information in the library. In library automation, there are limitations of data from users' activities because there are different characteristics of information that of amazon books selling e-commerce such as amazon, lazada or shopee. There are two main reasons for difference of technique between library system and other recommendations system. First, the users in the library system have different actions to do, thus users do only search on OPAC (Online public access catalog) [4] no review or any score to book. Second, the record of activities of the user only has a book's loan which cannot be enough data to analyze the data to calculate the recommender system model. Therefore, as we mentioned before for limited data for processing recommendation book, in this research we proposed the method to using a combined method applied for library recommendation systems.

To implement the recommendation system in library automation, it was necessary to apply the public interface of the user in the library main portal for search and retrieval of information in the OPAC module. In our scope, we develop open source library automation based on a small library [6] which has no more 2000 resources. For the small library in Thailand, it is defined as a library school which has more than 46000 places [7] in Thailand which can gain the benefit of our work. In addition, open source library automation was used widely in Thailand such as koha, Evergreen or openbiblio OBEC and WALAI AutoLib due to can be used without cost. In the field of recommendation systems of automation library, there are few researchers on automation library for example [8] using collaborative filtering based on the library loan records to recommend system.[10] also using single information to compute the users behaviors on a large scale book-loan logs of a university library. In another approach, [9] combine multi source data from book titles and loan logs with more accurate results for personalized recommendation.

In our studies, we combine based on multi source of information among titles, Dewey Decimal System [11] for book classification system and book-loan logs. The multiple data were implemented by support-vector [12] machines model in order to calculate similarity between the outlines in the BOOK Database. In our results, the three features were compared and conducted against three: loan logs, bibliographics similarity and combination of title. In this study, the information was employed from a small library in a private school which consists of a resource of material book 9654 bibliographics. Therefore, we proposed the recommendation system book based on combining multiple source data of title similarity bibliographic data, title, publisher and author. Finally, our research is structured as follows. Sections 1 and 2 are introduction and related work with some background on the principle of recommendation system on library automation. In Section 3 proposed approach to describe the algorithms and overview of our open source library automation. Section 4 is devoted to sum up the experiment result on our algorithms. In the end, Section 5 conclusions of this work and future work.

II. RELATED WORK

The research of book recommendation systems has been proposed rapidly due to the web technology adapt for digital library modernization. However, due to the context of library automation which has limited data from users to training in modern machine learning algorithms and different usage of behavior for using in recommendation system. Thus, few researches focused on implement the algorithms based on library load record, similarity of book titles or association rule method linking with book category. Many book recommendation systems have been developed such as LIBRA [26] used book information from amazon book data with applied text categorization to semi-structured data. K3Rec [27] was developed for k-3 readers. The idea was to find similar book content, book catalog suitability and other features to recommend book for children. Nova [28] used combine between collaborative filtering and content based filtering call hybrid method to find personalized book. Another researcher focus on book loan, [25][28] used FUCL mining technique based on association rule mining with linking of book loan information. The authors in [8][9][30] used book loan records from 39,442 users to combine between book loan information, association rule of book category and Matches/mismatches on Nippon Decimal Classification.[31] applies the information learned of an author-identification model by using convolutional neural network of book similarity.

III. PROPOSED METHODOLOGY FOR RECOMMENDATION SYSTEM

The purpose of this study is to develop the open source library automation implementation with a recommendation system which could suggest personalizing book favorites in order to increase user satisfaction and reduce the task of librarians. However, in this study we focus on algorithms in recommendation systems based on library automation only. Fig. 1 shows a brief detail of our open source library automation name as Angkaew autolib [13], the name of Angkaew refers to the name of a famous lake situated at the bottom of Suthep mountain in Chiang Mai University.

Angkaew autolib was develop based on core architecture of OpenBiblio the automated library system open source framework (<http://obiblio.sourceforge.net>) with consist of basic module for library automation such as OPAC, circulation, cataloging, and staff administration functionality. Angkaew autolib was develop by add on the feature of support Thai language, some user interface for librarian in Thailand and compatibility with modern programming with support php 7.0.

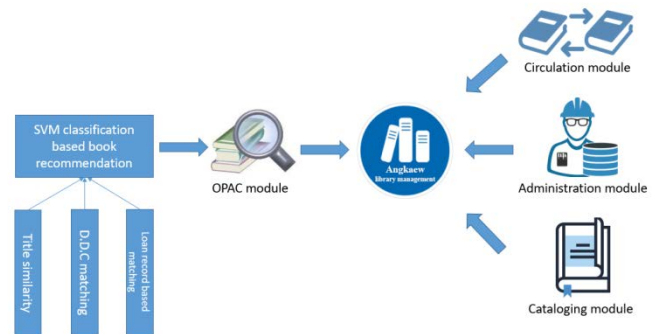


Fig. 1. The Overview of Angkaew Library Management.

A. The Character of School Library Users

The system aims for the school library and small library. However, as we mentioned earlier the recommendation in library automation may have different characteristics since users are mainly focused on academic purpose. For the academic purpose, the user may have searched for common courses among what they are actually learning in the class. For example, users borrow the practice English book because practice English book is a compulsory course in classroom. But the recommender should not represent all of English readings in the library but it recommends the related English in compulsory courses with popularity to adapt users' commonness. Another reason, there are not similar other commercial book selling, the library automation is no review data and no top rank selling in book-loan record. Even so the book-loan record can represent the interesting field of the user which can use this information to mapping the book category.

B. The Feature Selection of the Algorithms

Unlike the traditional collaborative filter [14][29] recommendation system, there is no scoring, popularity of item and user ratings data to calculate the preference towards related books. Thus, we proposed the method based on the book loan and multiple information to use as a feature for training in SVM classifier as follows.

1) *Similarity measures for book title*: The similarity of book title was used to determine the similarity between the titles which is the important feature to classify the related book. Thus, there are a large number of methods proposed to measure the similarity of the title matching [15]. However, the complexity of the structure of the title have varies for example a title can be phrase, a word or sentence with vary length and most importantly the most of methods may not be suitable for Thai languages because our languages have special character with upper/lower vowel [16]. Thai language is written from left to right without spaces between characters. Each character

has only one type that is no uppercase and lowercase of each character and combination with vowels to main consonant as shown in Fig. 2 and 3.

Consonants	44	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห พ อ ฮ
Vowels	18	ะ ั ำ ำ ิ ี ึ ื ึ ุ ู แ ำ ไ ใ ฤ ฦ
Tone marks	4	่ ้ ๊ ๋
Diacritics	5	์ ็ ๋ ๋ ๋
Numerals	10	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
Other symbols	6	๕ ๕ ๕ ๕ ๕ ๕
Total	87	

Fig. 2. Overview of Thai Language [16].



Fig. 3. Thai Word Combination of Upper Vowel and Lower Vowel.

Based on characters to represent in Thai language, we observed and enquired the librarian about the problem of information searching of automation library in the school of Banpasao Chiang Mai. We found the two main problems of book searching related to upper/lower vowels. First, the user always misinformation to spell the upper or lower vowel in the keyword search for example, *อยากให่เรื่องนี้ไม่มีโซคร้าย* (correct spell) vs *อยากไห่เรื่องนี้ไม่มีโซคร้าย* (miscorrect 3 upper voxel). In this case, the most similarity measures for title compute upper voxel as the main character to compute distance [17] which may not be accurate because the user is not aware the upper voxel is not important in Thai language and affects the meaning of the word in terms of computation. Second, the librarian cataloged the spell of wrong pronunciation in the title such as misspell upper and lower vowels, misspelling of transposition of word. In this case, we found misspell the word of title 43 from 2251 book title from the database record in school of Banpasao Chiang Mai. Thus, the title can always misspell pronunciation or user error input as follows.

คู่มือตัวสอบคณิตศาสตร์ CU - คู่มือตัวสอบคณิตศาสตร์ CU (2 misupper voxel).

คัมภีร์การเขียนภาษาอังกฤษธุรกิจ - คัมภีร์การเขียนภาษาอังกฤษธุรกิจ (1 misupper and 1 lower voxel).

หมายเหตุต้นกรุงรัตนโกสินทร์ - หมายเหตุต้นกรุงรัตนโกสินทร์ (1 misupper and 2 lower voxel)

กิจกรรมเฉลิมพระเกียรติ - กิจกรรมเฉลิมพระเกียรติ (1 transposition from keyword search).

Several methods of title similarity have been tested on our database the result of compare for measure shown in Table I.

TABLE I. COMPARE THE ALGORITHMS FOR STRING MEASURE

		Smith-Waterman	QGrams Distance	Levenshte in Distance	Modify Levenshte in Distance
คู่มือตัวสอบคณิตศาสตร์ CU	คู่มือตัวสอบคณิตศาสตร์ CU	92%	81%	92%	100%
คัมภีร์การเขียนภาษาอังกฤษธุรกิจ	คัมภีร์การเขียนภาษาอังกฤษธุรกิจ	94%	84%	94%	100%
หมายเหตุต้นกรุงรัตนโกสินทร์	หมายเหตุต้นกรุงรัตนโกสินทร์	96%	73%	89%	100%
กิจกรรมเฉลิมพระเกียรติ	กิจกรรมเฉลิมพระเกียรติ	91%	83%	91%	91%

In order to measure the book whose titles are similar to database, we modified the Damerau-Levenstein algorithm based on Levenstein's algorithm. We add the except deletion string to algorithms if the string equal upper or lower vowel in Thai language Unicode for more robust in term of common misspelling and suitable for user behavior in library automation system.

$$d_{a,b}(i,j) = \min \begin{cases} 0, & \text{if } i = j = 0 \\ 0, & \text{if } i \text{ or } j = \text{upper and lower vowel} \\ d_{a,b}(i-1,j) + 1, & \text{if } i > 0 \\ d_{a,b}(i,j-1) + 1, & \text{if } j > 0 \\ d_{a,b}(i-1,j-1), & \text{if } i, j > 0 \\ d_{a,b}(i-2,j-2) + 1, & \text{if } i, j > 1 \text{ and } |i| = |j| \text{ and } |i-1| = |j-1| \end{cases}$$

Where $+1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

- $d_{a,b}(i-1,j) + 1$ corresponds to a deletion (from a to b).
- $d_{a,b}(i,j-1) + 1$ corresponds to an insertion (from a to b).
- $d_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)}$ corresponds to match or mismatch, depending on whether the respective symbols are the same.

$d_{a,b}(i-2,j-2) + 1$ corresponds to a transposition between two successive symbols.

The DDC for systematic arrangement combination of association rule mining: As we mentioned earlier, the user searched the books which they are actually learning in the class or related the course. Based on this knowledge, the recommendation system should suggest the books are related to the same catalogue. For that reason, we consider the

catalogue of books to be an important feature to training in SVM. The DDC (Dewey decimal classification) number is the standard system for book classification systems used in Thai libraries both public and private schools. It used to organize and provide access to their book and other material collections in the library. Basically, there are the hierarchy of three levels of digit numbers representing the subject fields such as 400 for "Science" and 410 for "Linguistics". The first digit is the main field of the subject and the second consists of the sub-field. In our case, we decided to determine only the first digital for training in SVM because the only main subject field of book are flexible to recommend the book for example the user read "Improving your speaking English" the system may suggest the related field for languages such as Linguistics book or other language book the summaries of classification of DDC system shows in Table II. We summarise the number of books of classification based on DDC that have been loans in the database to combine association rule mining.

Unlikely traditional association rule mining [8], the concept of association rule is user borrows n books, x_i ($i = 1, \dots, n$) for once time, it call the set $\{x_1, \dots, x_n\}$ a "transaction". For instance, when a user borrows three books, A, B, and C, at one time and the A, B, C must be the same as the first digit of DDC, this transaction can be represented as $\{A, B, C\}$. In these transactions, we can provide the rule "the user who borrows book A also borrows book B. In addition, the user who borrows books B and C at one time also borrows book A. However, we add the rule in transaction must be the same as the first digit of DDC because to be more accurate for the behavior of the user in the library recommendation system is more likely to borrow in the same category of subject fields.

TABLE II. THE FIRST-LEVEL OF CLASSIFICATION OF DDC

D.D.C.	Subject fields	Number of Loans
000	Generalities	289
100	Philosophy	498
200	Religion	79
300	Social sciences	31
400	Language	1209
500	Science	879
600	Technology	801
700	Arts and recreation	112
800	Literature	370
900	History and geography	344

2) *Similarity measures for bibliographic information of book:* The Bibliographic book also benefits information when users prefer to find a similar book such as the book by the same author, nearly a year of publication, most viewed by other users. In addition, each bibliographic book has characteristics to extract the information. We extract the characteristics of a bibliographic for book recommendation.

a) *Date published of book:* The user tends to seek the book based on the date published of the book which they are interested in and borrowed. The system calculates the absolute

value based on the user selected booked between the book record publication dates. Thus, if the value is more less it mean it would be better to recommend the book.

b) *Book based category or Author:* Based on the investigation on the load data, the user tends to borrow the books of the same category or author which they borrowed before. Regarding to this knowledge, we calculated in the same way between logical 1 and 0 where 0 corresponds to matching with a same category or author and 1 corresponds to a no match.

c) *Number of views for the book:* The feature of Number of views for the book was calculated based on the views of books of the total number of users who search and click the detail on OPAC in the library automation system.

In order to combine each feature of (a)(b) and (c) we formulated an equation.

$$\text{Bib data} = w_1|y_s - y_r| + w_2(T_m + A_m) + w_3(N)$$

Where

w_i is weight of variable similarity.

$y_s - y_r$ is the different year of public.

T_m is Topic of book (book category).

A_m is Author.

N is number of view.

C. Book Recommendation System of Library Automation by SVM

Based on literature review, [18][19] used SVM as the machine learning method to book recommender systems as the effective tool. However, the effect of recommendation depends on the sources of information as the input to the learning process of SVM. As described before for each source of information, we implement the SVM learning which optimal parameters for classifying book into effective recommender system. We used three features as

- Similarity Measures for Book Title.
- The DDC for systematic arrangement combination of Association Rule Mining.
- Similarity Measures for Bibliographic Information of book.

We implement the module of recommender system with LibSVM [20][21] version 3.24 on our open source automated library system. The libSVM supports vector classification with "C-Support Vector Classification". The system showed the four highest scores of the book which combine based on three features that we mention before and recommended the book arrangement in OPAC system. As we mentioned before, we developed an open source library automation name as Angkaew with implement the recommendation feature as in figure x can be download in Department of library and information science Faculty of humanities Chiang Mai university (<http://lis.human.cmu.ac.th/>). The book recommendation system was included below the OPAC feature. The example used the keyword of "English languages

in grade 5”, the algorithms arrange the top four highest scores to show from highest to lower respectively the program shown in Fig. 4.

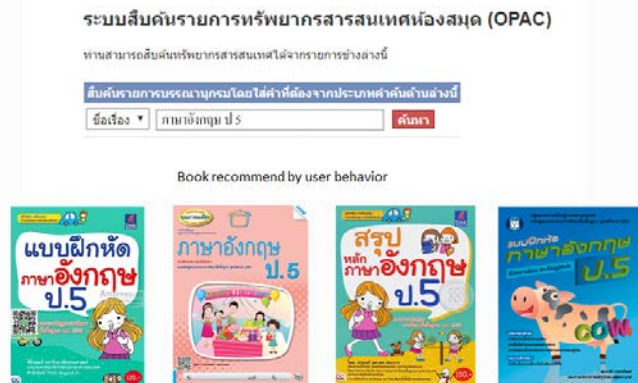


Fig. 4. Example of OPAC Implement with Book Recommendation.

IV. EXPERIMENT RESULT

In order to experiment the effectiveness of book recommendation, two types of evaluation were employed in this research: qualitative and quantitative data.

A. Participate

Our participants were thirty students of lower-secondary and thirty-two upper-secondary in the school of Banpasao Chiang Mai which have the similar background and have experience to use OPAC service to search the book in library. The participants were registered in the system and have loan records in the library automation at least once per month.

B. Qualitative Data

To evaluate the book recommendation in terms of qualitative data, we asked the participants to describe their satisfaction based on level of interest in each book (during using the OPAC with implement the recommendation system) using the following five-point scale of interest which is similar to use by [8][22]. 2: Very interested 1: Interested 0: Not interested x: book is not related in my mind A: Already borrowed or have read before. For each participant searched the book on OPAC for ten times the result shown as Table III.

Based on the satisfaction questionnaire, our recommendation system can persuade the book recommender of “Very interested” and “interested” by 14.5% and 22.5% respectively. To better understand, we also interview the students reported that they feel language, science and technology catalogue are the best recommendation book. However, the negative result are “Not interested” and “Book is not related in my mind” by 31.6% and 19.1% respectively. For this result, the students reported even the book is the same catalogue which they prefer but the titles are not related to their interests. Finally, the mark of “Already borrowed or have read before” of 12.0%. To investigate this result, the students reported that they read it before to use the system but there is no case of “already borrowed” in the same library automation system.

C. Qualitative Data Analysis

Data analysis of OPAC usage: We are also monitoring the user behavior based on usage of OPAC [23] activity by library. The circulation records and library autolib log-file were collected between July 2019 and December 2019 with support by the administrator of the school of Banpasao Chiang Mai. At the end of each month, the data were recorded from the Angkeaw autolib server. The aim of log-file was to data analytic of library collection usage for recommendation system in OPAC module.

Log-files were collected into a database in order to verify the effectiveness of our algorithms for recommendation system. The database of activity provides the information for user’s last activity in OPAC module such as user using keyword to search in OPAC, click on book recommendation. However our database cannot analyze whether the user pay attention or impressive to the book recommendation on it. We compared between two groups of usage. Group of “search keyword on OPAC with click recommender” defined as users who use keyword to search in OPAC module and click on book recommendation system per time of using one keyword and the group of “search keyword on OPAC without click recommender” defined as users who use keyword to search in OPAC module without any click on recommendation system per time of using one keyword. Based on our statistic, the frequency of using OPAC is vary depending on month. To analyze data, it have highest on August (52 per day) and November (56 per day) due to midterm and Final exam. We represented the statistic show as activity per month. The statistic is shown in Fig. 5.

TABLE III. THE FIRST-LEVEL OF CLASSIFICATION OF DDC

	Lower-secondary students (300 times)	Upper-secondary students (320 times)	Total (620 times)
Very interested	36 (12%)	54 (16.8%)	90 (14.5%)
interested	77 (25.6%)	63 (19.6%)	140 (22.5%)
Not interested	94 (31.3%)	102 (31.8%)	196 (31.6%)
Book is not related in my mind	52 (17.3%)	67 (20.9%)	119 (19.1%)
Already borrowed or have read before	41 (13.6%)	34 (10.6%)	75 (12.0%)

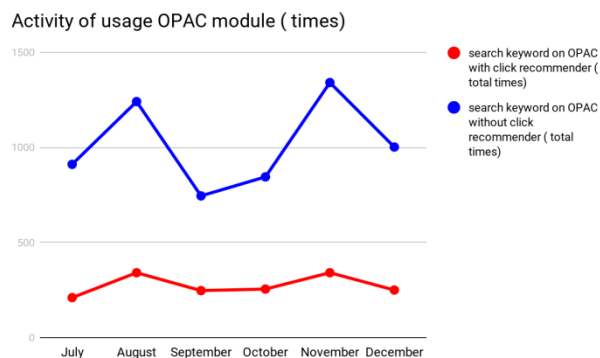


Fig. 5. Show Activity of usage OPAC Module in our System.

1) *Experiment and data analysis with precision:* In order to measure the accuracy of our algorithms, we employed the evaluation metrics of precision [24] and have been used for book recommendation [10]. The evaluation metrics of precision is defined as formula 1. The recommendation books predicted in this experiment refer to the book categories based on Dewey Decimal Classification. We evaluate the accuracy of algorithms by the average precision of each user which divides the performance of each category of Dewey Decimal Classification.

$$Precision = \frac{\text{number of hit books}}{\text{number of recommended books}}$$

We collected experiment data by the monitoring system of the user from the total 431 users in Angkaew Autolib in July 2019 - December 2019. However, our algorithms compute based on DDC category for systematic arrangement of book recommendation. The reason why our algorithm computes and recommends only the same DDC category of user view is because users in the library always borrow the books of the same category unlike commercial book stores. Thus, we measured the precision of each DDC category listed in Table IV.

TABLE IV. COMPARE THE PERFORMANCE OF PRECISION TO DDC CATEGORY

	000 (Generalities)	100 (Philosophy)	200 (Religion)	300 (Social sciences)	400 (Language)	500 (Science)	600 (Technology)	700 (Arts and recreation)	800 (Literature)
P	0.05	0.13	0.21	0.18	0.31	0.27	0.25	0.11	0.07

The precision of each DDC category is shows in Table IV, where p denotes the precision of our algorithms which the average precision of algorithms is 0.164. In addition, the top three highest precision of DDC categories are Language (0.31), Science (0.27) and Technology (0.25).

V. CONCLUSIONS AND DISCUSSION

We proposed the module of Automation library using a book recommendation system by a support vector machine implemented with three features of information: similarity measures for book title, DDC for systematic arrangement of materials and similarity measures for bibliographic information of book. Based on experiment result, our algorithm improved usage of user in library automation from both qualitative and quantitative data. For qualitative data based on satisfaction questionnaire, the results of our recommendation system show user feel very interested (14.5%) through usage of OPAC module. In addition, the result of recorded user behavior based on usage activity showed effective 1644(21.27%) from 7729 times using OPAC click view on book recommendation system. However, we designed the algorithms based on usage of student behavior with adapt on OPAC module in library automation. Our algorithm suitable for small library such as library in school, small organization and specialized library with book resource less than 10000 books and not too much data information of loan record.

An algorithm specifically designed based on user behavior using library automation using a support vector machine and specific feature selection of modification similarity measures of book title in Thai language and multiple bibliography book information. However, the limitation of algorithms fit for small dataset of book resource because due to computation time for whole bibliographic book for each time. Our future work will focus on improving the algorithms of Title similarity specifically on Thai language for the whole category of book with implementation on natural language processing to understand the nature of each title book in order to provide the personalized book.

VI. REFERENCES

- [1] F. Ricci, L. Rokach and B. Shapira, "Introduction to recommender systems handbook," In Recommender Systems Handbook. Boston, MA.: Springer, 2011, pp. 1-35.
- [2] V. Adamson, JISC & SCONUL Library Management Systems Study PDF (1 MB). Sheffield, UK: Sero Consulting, 2008, p. 51.
- [3] T. R. Kohtanek, "1-The Evolution of LIS and Enabling Technologies," In Library Information Systems: From Library Automation to Distributed Information Access Solutions. Westport, CT: Libraries Unlimited, 2002, p. 5.
- [4] C. L. Borgman, "Why are online catalogs still hard to use?," Journal of the American society for information science, vol. 47, no. 7, pp. 493-503, 1996.
- [5] N. P. Mahwasane and N. P. Mudzielwana, "Challenges of Students in Accessing Information in the Library: A Brief Review", Journal of Communication, vol. 7, no. 2, pp. 216-221, 2016.
- [6] V. T. Kamble, H. Hans Raj and S. Sangeeta, "Open Source Library Management and Digital Library Software", DESIDOC Journal of Library & Information Technology, vol. 32, no. 5, pp. 388-392, 2012.
- [7] D. Rhein, "International Higher Education in Thailand: Challenges within a Changing Context", Journal of Alternative Perspectives in the Social Sciences, vol. 8, no. 3, 2017.
- [8] K. Tsuji et al., "Use of library loan records for book recommendation", in 2012 IIAI International Conference on Advanced Applied Informatics, 2012, pp. 30-35.
- [9] K. Tsuji et al., "Book Recommendation Based on Library Loan Records and Bibliographic Information", Procedia - Social and Behavioral Sciences, vol. 147, pp. 478-486, 2014.
- [10] C. Chen et al., "Book recommendation based on book-loan logs", In International Conference on Asian Digital Libraries, Springer, Berlin, Heidelberg, 2012, pp. 269-278.
- [11] M. Dewey, A classification and subject index for cataloguing and arranging the books and pamphlets of a library. Brick row book shop, Incorporated, 1876.
- [12] C. Cortes and V. Vapnik, "Support-vector networks", Machine learning, vol. 20, no. 3, pp. 273-297, 1995.
- [13] K. Puritat and K. Intawong, "Development of an Open Source Automated Library System with Book Recommendation System for Small Libraries", in 2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), 2020, pp. 128-132.
- [14] L. Terveen and W. Hill, "Beyond recommender systems: Helping people help each other", HCI in the New Millennium, vol. 1, no. 2001, pp. 487-509, 2001.
- [15] N. Gali, R. Mariescu-Istodor and P. Fränti, "Similarity measures for title matching", In 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1548-1553.
- [16] H. T. Koanantakool, T. Karoonboonyanan, and C. Wutiwathchai, "Computers and the Thai language", IEEE Annals of the History of Computing, vol. 31, no. 1, pp. 46-61, 2009.
- [17] N. Gali et al., "Framework for syntactic string similarity measures", Expert Systems with Applications, vol.129, pp. 169-185, 2019.

- [18] K. Tsuji, "Book Recommender System for Wikipedia Article Readers in a University Library", In 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), 2019, pp. 121-126.
- [19] K. Anwar, J. Siddiqui, and S. S. Sohail, "Machine learning-based book recommender system: a survey and new perspectives", International Journal of Intelligent Information and Database Systems, vol. 13, no. 2-4, pp. 231-248, 2020.
- [20] C. C. Chang, and C. J. Lin, "LIBSVM: A library for support vector machines", ACM transactions on intelligent systems and technology (TIST), vol. 2, no. 3, pp. 1-27, 2011.
- [21] C. C. Chang, and C. J. Lin, C.-J.: LIBSVM: a library for support vector machines, 2004.
- [22] T. Harada, and K. Masuda, "A trial approach of weighting for library loan records for developing a book recommendation system", In Digital Libraries, vol. 38, pp. 54-66, 2010.
- [23] H. L. Chen, and B. Albee, "An open source library system and public library users: Finding and using library collections", Library & information science research, vol. 34, no. 3, pp. 220-227, 2012.
- [24] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", 2011.
- [25] P. Jomsri, "FUCL mining technique for book recommender system in library service", Procedia Manufacturing, vol. 22, pp. 550-557, 2018.
- [26] R. J. Mooney, P. N. Bennett, and L. Roy "Book recommending using text categorization with extracted information", In Proc. Recommender Systems Papers from 1998 Workshop, Technical Report WS-98-08, vol. 1188, 1998.
- [27] M. S. Pera, and Y. K. Ng, "Analyzing book-related features to recommend books for emergent readers", In Proceedings of the 26th acm conference on hypertext & social media, 2015, pp. 221-230.
- [28] D. Pathak, S. Matharia, and C. N. S. Murthy, "NOVA: Hybrid book recommendation engine", In 2013 3rd IEEE International Advance Computing Conference (IACC), 2013, pp. 977-982.
- [29] L. Xin, "Collaborative book recommendation based on readers' borrowing records", In 2013 International Conference on Advanced Cloud and Big Data, 2013, pp. 159-163.
- [30] K. Tsuji, et al., "Book recommendation using machine learning methods based on library loan records and bibliographic information", In 2014 IIAI 3rd International Conference on Advanced Applied Informatics, 2014, pp. 76-79.
- [31] H. Alharthi, D. Inkpen, and S. Szpakowicz, "Authorship identification for literary book recommendations", In Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 390-400.

Econometric Analysis of Stock Market Performance during COVID-19 Pandemic: A Case Study of Uzbekistan Stock Market

Uzbekistan Stock Market Performance during COVID-19 Pandemic

Mansur Eshov¹, Walid Osamy², Ahmed Aziz³, Ahmed M. Khedr⁴

Tashkent State University of Economics, Tashkent, Uzbekistan¹

Computer Science Department, Faculty of Computers and Artificial intelligence, University of Benha, Egypt²

Tashkent State University of Economics, Tashkent, Uzbekistan³

Computer Science Department, Faculty of Computers and Artificial intelligence, University of Benha, Egypt³

Computer Science Department, University of Sharjah, Sharjah 27272, UAE⁴

Abstract—This article highlights the impact of the Coronavirus disease (COVID-19). COVID-19 pandemic on the stock market of Uzbekistan on the basis of empirical research and the main factors affecting the stock market are identified as well. Secondary statistical data were collected from the Tashkent Stock Exchange, the Central Bank of the Republic of Uzbekistan, the State Statistics Committee of the Republic of Uzbekistan and other public funds, and the regression equation of the SEM-model of the impact of the Covid-19 pandemic on the stock market of Uzbekistan was formed. In particular, indicators such as the latest daily and total number of people infected with Covid-19 in the Republic of Uzbekistan, the total number of recovered people after being infected with Covid-19, the total number of people who died of the disease, the daily number of recovered people post-infection, the stock market index of Uzbekistan, Uzbekistan Indicators such as the number of daily securities traded on the Republican Stock Exchange "Tashkent", the exchange rate of the US dollar set by the Central Bank of the Republic of Uzbekistan were selected as the main factors. The constructed regression equation was examined using F-statistics, Student's t-test, and multicorrelation tests to determine the level of adequacy. The authors identify factors based on a systematic analysis of the scientific work of world-renowned scientists on major stock markets and creates a SEM-model of the factors affecting the Uzbek stock market during the pandemic.

Keywords—Stock market; factors; SEM-model; COVID-19; global economy

I. INTRODUCTION

Although the COVID-19 pandemic has only recently begun, the coronavirus, a new challenge and test for humanity, has also had a strong negative impact on the economies of the world's poorest countries. Even though there is no clear evidence and proof of the biological origin of Covid-19 coronavirus, its positive and negative effects on the world economy are recognized by many scientists and publishers.

The stock market is a barometer of the economy, and the capital market reflects the general state of the country's economy to some extent. Increasing uncertainty in the stock markets leads to sharp fluctuations in prices. Under the

influence of the pandemic, unprecedented changes have taken place in world stock markets. As a result, risks in the global stock market against pandemics have increased significantly, and the uncertainty of the disease and the associated economic losses have made markets highly volatile and unpredictable.

Therefore, along with other field scientists, economists have also conducted research on the impact of the pandemic on the economy, particularly the stock market. Recently, global scientists have developed various forecast scenarios on the impact of the pandemic on the stock market. Research is currently trying to answer the following questions: How can risk in stock markets respond to a pandemic? Are systemic risks increasing worldwide? What are the potential effects of policy intervention?

According to statistics, in the context of the pandemic, along with the collapse of the U.S. stock market, stock markets in Europe and Asia were also disrupted. The highest decline in global stock market indices occurred in March. The Dow Jones industrial average fell by 17.3%, the S&P 500 fell by 14.98% and the Nasdaq fell by 12.64%. The Financial Times Stock Exchange Index (FTSE [19]), a key indicator of the UK stock market, fell to its worst level since 1987 in March 2020, that is by 16.99%, while the stock market in Japan fell by an average of 6.4% a year [18].

During the pandemic, stock prices also fell sharply in line with indices in major stock markets. In the U.S., the market price of shares, which includes the capitalization level of all companies, fell by 33.1%, [20] and in Europe by 39.95%. The situation in the Pacific region was better than in the US and Europe, with a decline in the market value of shares in the region made up 26.89% [21].

Thus, the development and implementation of scientific proposals aimed at mitigating the impact of the ongoing COVID-19 coronavirus pandemic on the economy, including the stock markets of countries around the world, is highly relevant today.

The COVID-19 crisis poses a significant risk to Uzbekistan's ambitious economic and social transition. The

global COVID-19 pandemic is the most severe crisis Uzbekistan has faced since the economy's recovery from the breakup of the Soviet Union. It has adversely affected the domestic economy and resulted in declines in employment, well-being, and incomes. Growth is projected to slow to 1.5 percent in 2020, while lower exports and remittances are expected to widen the current account deficit to almost 10 percent of Gross domestic product (GDP). Addressing the external shock and the domestic impact of COVID-19 is expected to require additional external financing of about US\$ 4 billion (7 percent of GDP). The COVID-19 crisis has almost entirely extinguished GDP growth in 2020. To mitigate the economic, social, and health consequences of the pandemic, the Government has been taking anti-crisis policy measures. Persistent COVID-19 disruptions at the local and international levels have tempered prospects for a quick recovery in 2021. Nevertheless, Uzbekistan's outlook remains positive as reforms continue to shift the economy toward greater resource efficiency and private sector growth. This work highlights the impact of the COVID-19 (coronavirus) pandemic on the Uzbekistan stock market based on an empirical research. The main factors affecting the stock market are identified as well.

The paper is further structured as follows: Section II provides the literature review and in Section III, presents the Research methodology. The Analysis and results are included in Section IV and finally Section V concludes this research paper.

II. LITERATURE REVIEW

The outbreak of the COVID-19 pandemic in the world has had a negative impact on the global economy, leading to a sudden cessation of supply and demand for a period of time, with industrial enterprises, most service sectors and a large proportion of small businesses stalling for some time. This, in turn, has negatively affected major stock markets such as the Dow Jones, Nasdaq, S&P 500. A number of scientific studies have been conducted by the world community of scientists to study these problems.

The impact of various infections, epidemics and pandemics on the global economy in the 21st century has been studied by [7]. (2020), and the laws governing the impact of pandemics on the economy have been identified.

In his study, [11] examined the reaction of stock markets to the COVID-19 pandemic. According to the author, the growth of confirmed COVID-19 cases had a negative impact on stock markets, using data on daily COVID-19 approved cases and deaths, and data on stock market earnings from 64 countries from 22 January 2020 to 17 April 2020. That is, as the number of approved cases increases, the profitability of the stock market decreases. The author also notes that stock markets have a greater impact on the increase in confirmed cases than deaths [11].

The team of authors in [12] studied the impact of COVID-19 on the Indian financial market and compared it with the results of the last two structural changes in the country's economy: demonetization and goods and services introduction of the goods and services tax (GST). Daily stock yields from January 3, 2003 to April 20, 2020 will be negative for all

indices during the COVID-19 spread, in contrast to the post-demonetization and post-GST phases based on net foreign institutional investors and exchange rate [17] data. According to the study, the COVID-19 pandemic will have a significant impact on stock yields relative to demonetization and GST performance in India. Foreign researchers Kerstin Lopatta, Kenji Alexander, Laura Gastone and Thomas [13] used data from nearly 300 international companies included in the leading stock market indices of ten countries to report on companies' reporting practices during the COVID-19 coronavirus pandemic studied the impact on capital values and the assessment of capital market risks. The authors draw two main conclusions from the research findings. First, it was found that beta-value declines in the annual reports of firms reporting the coronavirus crisis using the capital market model.

Second, comparing the earnings before and after the publication of the annual report, it was found that the earnings of firms that report coronavirus pandemics in their annual reports have improved significantly. The authors' research shows that investors value the transparency of information about companies and their ability to promptly make global changes in the reporting process. Spanish researchers have studied market variability (volatility) and stock market comparisons during the COVID-19 pandemic and formed a number of conclusions. In particular, you should not buy or sell based on news or daily market movements. Fear, speculation and uncertainty increase volatility in the market [15][16].

The impact of the Covid-19 pandemic on the stock market in the United States and the problems of its integral link with geopolitical and political uncertainties have been studied by Arshian Sharifa, Chaker Aloui, Larisa [1].

The influence of the stock market reaction to the pandemic in the first quarter of the onset of the Covid-19 pandemic was analyzed by a number of scientists [2] preliminary results were obtained [4-5]. These studies have shown an increase in the share prices of food companies and IT software companies (Zoom, Yandex, Vine, etc.) in the stock market.

Complex changes in the stock market, changes in the economies of the U.S. and European countries, according to the statements of the World Health Organization and the conclusions on the COVID-19 and studied in depth by other scholars [6-8].

Based on the recommendations of medical scientists on face and hand masks, the popularity of cashless payments has led to an increase in the price of digital money such as Bitcoin in the world and its impact on the stock market has been studied by [3].

The work in [9] studied the stock markets of 12 countries with high liquidity and proved that people's panic increases market risk and negatively affects stock prices.

Recent research by [10] has shown that the use of gold in stock market risk hedging is advisable. In the complex changes of the stock market at different time intervals of the pandemic, it was found that only the stability of gold is an acceptable tool in preventing financial risks.

The analysis of the above literature shows that the impact of the Covid-19 coronavirus pandemic on the stock markets, which is one of the most important financial institutions in developed and rapidly developing countries, has been thoroughly studied. This, in turn, indicates that stock markets make a significant contribution to the economic development of these countries. However, Uzbek economists have not yet done enough research on the impact of the pandemic on the Uzbekistan stock market.

III. RESEARCH METHODOLOGY

In the course of the study, the direct and indirect factors affecting the stock market were selected, the relationship between them was made on the basis of correlation analysis, and the regression equation based on the Structural equation Model (SEM-model) of the impact of the Covid-19 pandemic on the Uzbek stock market. Secondary statistics for the econometric model were collected from the Tashkent Stock Exchange, the Central Bank of the Republic of Uzbekistan, the State Statistics Committee of the Republic of Uzbekistan and other open sources.

IV. ANALYSIS AND RESULTS

In late December 2019, China officially reported to the World Health Organization that a coronavirus infection COVID-19 caused by an unknown type of pneumonia had occurred in Wuhan Province. The virus was detected to be transmitted from person to person, and there was no cure for it when the epidemic began.

In January 2020, the first cases appeared outside of China. On March 13, the number of cases exceeded 132,000, with 4,947 deaths, compared to 123 in the countries where the virus was detected.

Against the backdrop of a new disease outbreak, investors have reconsidered their views on the future of the global economy. Restrictive measures introduced in different countries have negatively affected almost all sectors related to consumer activities: tourism, trade, catering, entertainment and others. Under quarantine, people spend less money and move on.

Traders began to get rid of shares of airlines, oil companies, home appliance manufacturers and other companies in anticipation of declining profits and earnings. Indices of the world's leading stock exchanges fell. Italy's FTSE market index for the Borsa Italiana (MIB) index alone lost 29.8% from February 19 to March 11. Interestingly, the Chinese stock market began to recover slowly, while in other countries it continued to decline.

The COVID-19 coronavirus affects 218 countries and territories around the world and 2 international vehicles. The list of countries and regions and their continental regional classification is based on the United Nations Geoscheme [14] (Table I).

Asian stock exchanges have suffered the most from the coronavirus. In particular, the KOSPI 200 index, which consists of shares of the 200 largest companies in South Korea, fell to a 10-year low.

European stock exchanges in France, the United Kingdom and Germany also suffered from the effects of the coronavirus pandemic.

Measures are being taken around the world to mitigate the effects of the coronavirus pandemic. A number of programs are being developed and implemented to stabilize the impact of the coronavirus pandemic on the global economy and financial markets.

Many countries have already suffered from supply chain disruptions as China is a supplier of many types of components, including electronics, automotive, machinery and various equipments for such industries.

As for Uzbekistan, according to 2019 data, it is in the structure of imports. China is the largest partner (\$ 5.1 billion), followed by Russia (\$ 4.1 billion) and Korea (\$ 2.6 billion). It is important to pay attention to the composition of imports - China and Korea are the main suppliers of machinery and equipment to Uzbekistan, supplying about 50% of the total.

According to the Central Bank of Uzbekistan for 9 months of 2019, remittances (personal remittances) to the Republic of Uzbekistan amounted to 4.5 billion US dollars, of which 85% from Russia and 6% from Kazakhstan. While maintaining current energy prices, remittances to Uzbekistan expect to fall by 15-20 percent in 2020. In times of crisis, investors begin to sell valuable assets such as securities and invest in defense assets. Protective assets include U.S. government bonds and gold.

This could be a positive result, given that Uzbekistan exported \$ 4.9 billion worth of gold, or 27.5 percent of total exports. While maintaining the current price range for 2020 and the volume of exports in tons, this figure could exceed 25%.

In the econometric analysis of the factors affecting the development of the stock market of Uzbekistan COVID-19 coronavirus pandemic, first of all, it is necessary to choose the theoretically and logically correct factors. To study the impact of the Covid-19 coronavirus pandemic on the development of the Uzbek stock market the following is considered (Fig. 1):

- 1) The daily turnover of the Covid-19 coronavirus infection in the Republic of Uzbekistan -x1,
- 2) Total number of patients-x2,
- 3) Total number of patients with coronavirus Covid-19 in the Republic of Uzbekistan -x3,
- 4) Total number of deaths from coronavirus infection in the Republic of Uzbekistan -x4,
- 5) Days of recovery from coronavirus infection in the Republic of Uzbekistan Covid-19 -x5,
- 6) Stock Market Index of the Republic of Uzbekistan -x6,
- 7) Number of daily securities traded on the Republican Stock Exchange "Tashkent" of the Republic of Uzbekistan -x7,
- 8) US dollar exchange rate set by the Central Bank of the Republic of Uzbekistan -x8. Statistics on the number of cases, cures and deaths of Covid-19 coronavirus infection in the Republic of Uzbekistan are available on the official website [15].

Based on the selected factors, it is expedient to determine their degree of correlation by means of a correlation coefficient in a special program Stata 16 (Table II).

If the descriptive statistics examined of the sample, it noticed that the number of observations is 181 and the factors are 8, and the mean deviations and boundary values are shown in Table IV.

According to the table, there is a strong correlation between the resulting factor and the selected factors, and since the correlation between the factors is dense and the conditions $|r_{(x_1, x_2)}| < 0.8$, there is no multicollinearity between the factors and regression equation can be constructed. The regression equation shows what functional relationship exists between the resulting factor and the selected factors (Table III).

TABLE I. WORLD COVID-19 CORONAVIRUS INFORMATION AS OF DECEMBER 10, 2020 (2020)

№	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/1M pop	Deaths/1M pop	Total Tests	Tests/1M pop	Population
	World	69,332,094	116.481	1,577,777	2.885	48,088,704	19,665,613	106.846	8.895	202.4			
1	USA	15,822,734	+1,371	296.745	+30	9,231,811	6,294,178	27.329	47.679	894	231,115,488	642.185	331,859,797
2	India	9,767,371	+5,045	141.772	+37	9,253,306	372.293	8.944	7.047	102	150,759,726	108.775	1,385,975,035
3	Brazil	6,730,118		179.032		5,901,511	649.575	8.318	31.563	840	25,700,000	120.529	213,225,942
4	Russia	2,569,126	+27,927	45.28	+562	2,033,669	490.177	2.3	17.601	310	81,021,364	555.084	145,962,223
5	France	2,324,216		56.648		173.247	2,094,321	3.041	35.572	867	28,023,593	428.905	65,337,492
6	Italy	1,770,149		61.739		997.895	710.515	3.32	29.296	1.022	23,504,588	389.005	60,422,323
7	UK	1,766,819		62.566		N/A	N/A	1.272	25.966	919	46,344,703	681.103	68,043,572
8	Spain	1,725,473		47.019		N/A	N/A	2.179	36.898	1.005	24,101,272	515.394	46,762,812
9	Argentina	1,475,222		40.222		1,311,488	123.512	3.688	32.509	886	4,145,226	91.348	45,378,388
10	Colombia	1,392,133		38.308		1,287,597	66.228	2.376	27.232	749	6,855,035	134.093	51,121,380
11	Germany	1,242,253		20.704		902,100	319.449	4.278	14.806	247	30,494,036	363.446	83,902,518
12	Mexico	1,205,229	+11,974	111.655	+781	889.168	204.406	3.515	9.305	862	3,086,510	23.829	129,528,514
13	Poland	1,102,096	+13,749	21.630	+470	792.119	288.347	1.775	29.134	572	6,586,361	174.112	37,828,199
...													
76	Libya	89.183	+661	1.273	+12	59,222	28,688		12,902	184	474,223	68,606	6,912,270
77	Bahrain	88.495		347		86,518	1,630	6	51,243	201	2,149,366	1,244,578	1,726,984
78	China	86.673	+12	4.634		81,754	285	5	60	3	160,000,000	111,163	1,439,323,776
79	Lithuania	83.883	+3,330	735	+31	34,975	48,173	173	31,008	272	1,372,990	507,529	2,705,247
80	Kyrgyzstan	76.391	+379	1.306	+3	68,894	6,191	106	11,625	199	533,736	81,223	6,571,274
81	Malaysia	76.265		393		65,124	10,748	127	2,343	12	2,851,220	87,600	32,547,996
82	Ireland	74.900		2.102		23,364	49,434	38	15,095	424	2,063,450	415,848	4,962,029
83	Uzbekistan	74.498	+146	611		71,740	2,147	179	2,212	18	1,377,915	40,910	33,681,756

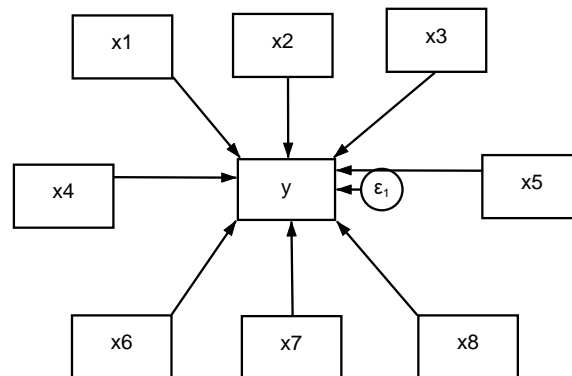


Fig. 1. Conceptual view of the Multifactorial SEM-model of the Impact of the Covid-19 Coronavirus Pandemic on the Development of the Stock Market in the Republic of Uzbekistan.

TABLE II. ILLUSTRATIVE STATISTICS OF FACTORS INFLUENCING THE DEVELOPMENT OF THE STOCK MARKET IN THE CONDITIONS OF THE PANDEMIC OF COVID-19 VIRUS IN THE REPUBLIC OF UZBEKISTAN

Variable Name	Vars: 9		
	Storage Type	Display Format	Variable Label
y	float	%8.0g	y
X ₁	long	%8.0g	X ₁
X ₂	int	%8.0g	X ₂
X ₃	long	%8.0g	X ₃
X ₄	int	%8.0g	X ₄
X ₅	int	%8.0g	X ₅
X ₆	float	%8.0g	X ₆
X ₇	long	%8.0g	X ₇
X ₈	float	%8.0g	X ₈

TABLE III. STANDARD DEVIATION VALUES OF FACTORS INFLUENCING THE DEVELOPMENT OF THE STOCK MARKET IN THE CONDITIONS OF THE PANDEMIC OF COVID-19 VIRUS IN THE REPUBLIC OF UZBEKISTAN

Variable	obs	mean	Std.Dev.	min	max
y	181	430343.9	1055343	1703.7	6857460
X ₁	181	28024.02	26597.18	8	73094
X ₂	181	368.3094	428.1241	3	3030
X ₃	181	25223.78	25955.46	0	70337
X ₄	181	216.3039	229.979	0	610
X ₅	181	371.9834	524.7122	0	3582
X ₆	181	622.8969	10.2359	598.56	647.08
X ₇	181	2243300	1.34e+07	5533	1.75e+08
X ₈	181	10325.6	1301.471	9520.3	20225

TABLE IV. CORRELATION ANALYSIS OF THE IMPACT OF THE COVID-19 VIRUS PANDEMIC ON THE DEVELOPMENT OF THE STOCK MARKET IN THE REPUBLIC OF UZBEKISTAN

Variable	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
X ₁	1.0000	-	-	-	-	-	-
X ₂	0.3714	1.000					
X ₃	0.9963	0.3245	1.000				
X ₄	0.9956	0.3229	0.9987	1.0000			
X ₅	0.3640	0.9397	0.3241	0.3171	1.000		
X ₆	0.1944	0.2074	0.1514	0.1357	0.1920	1.000	
X ₇	0.0746	0.0123	0.0775	0.0761	-0.0066	-0.014	1.000
X ₈	0.1106	0.2800	0.0761	0.0830	0.2558	0.2195	0.0078

It is advisable to use the modern Stata 14 or Eviews program to construct the regression equation.

According to Table III, the significance of the correlation should be examined against certain criteria. For example, the new daily number of Covid-19 coronavirus infections in the Republic of Uzbekistan is closely related to the total number of Covid-19 coronavirus deaths in the Republic of Uzbekistan ($|r_{([x_1, x_4])}| = 0,9956$), in fact, it is logical that the number of deaths due to the disease is directly related to the number of people suffering from the disease. A similar situation is closely related to the total number of coronavirus infections in the Republic of Uzbekistan ($|r_{([x_5, x_2])}| = 0.9397$). However, there is a strong correlation between the

total number of people cured of Covid-19 coronavirus infection in the Republic of Uzbekistan -X₃ and the total number of deaths from Covid-19 coronavirus infection in the Republic of Uzbekistan -X₄, which requires the exclusion of one of these factors.

To test the hypothesis that the development of the stock market in the Republic of Uzbekistan is affected by the Covid-19 coronavirus pandemic, a regression model is constructed between immutable and unchanging.

A regression matrix was obtained on the basis of a two-factor SEM-model of the impact of the Covid-19 coronavirus pandemic on the development of the stock market in the Republic of Uzbekistan (Table V).

TABLE V. MULTIVARIATE REGRESSION MATRIX OF THE IMPACT OF THE COVID-19 CORONAVIRUS PANDEMIC ON THE DEVELOPMENT OF THE STOCK MARKET IN THE REPUBLIC OF UZBEKISTAN

Source	SS	df	MS	Number of obs = 181 F(8, 171) = 1.66 Prob > f = 0.1107 R-squared = 0.0718 Adj R-Squared = 0.0286 Root MSE = 1.0e+06		
Model	1.4390e+13	8	1.7988e+12			
Residual	1.8608e+14	172	1.0819e+12			
Total	2.0047e+14	180	1.1137e+12			
y	Coef	Std.Err.	t	P > t	95% conf	interval
X ₁	73.13249	55.87447	1.31	0.192	-37.15544	183.4204
X ₂	-395.9719	592.4918	-0.67	0.505	-1565.463	773.5194
X ₃	-15.77854	73.09006	-0.22	0.829	-160.0475	128.4904
X ₄	-5804.004	7961.744	-0.73	0.467	-21519.31	9911.301
X ₅	19.93819	461.8393	0.04	0.966	-891.6643	931.5407
X ₆	-3252.676	10417.18	-0.31	0.755	-23814.64	17309.29
X ₇	0148077	0.005811	2.55	0.012	0.0033376	0.0262777
X ₈	-19.45844	66.03357	-0.29	0.769	-149.7989	110.8821
_cons	2366508	6375758	0.37	0.711	-1.02e+07	1.50e+07

The reliability and adequacy of the regression equation should be checked against the criteria. The inspection was performed using the Stata 14 program. The Fisher criterion is F_stat = 0.11, which is greater than 0.05, which means that the model is not statistically significant. The p-value of each factor in the regression equation of factors x1, x2, x3, x4, x5, x6, and x8 is 0.19, 0.50, 0.82, 0.46, 0.96, 0.75, 0.76, respectively, and no factor other than x7 does not matter.

According to the table, there is a strong correlation between the resulting factor and the selected factors, and since the correlation between the factors is dense and the conditions |r_([x_(1), x]_2) | < 0.8 -salt of the regression equation on the basis of the model, the functional relationship is studied between the resultant factor of the regression equation and the selected factors (Table VI).

TABLE VI. DOUBLE CORRELATION OF FACTORS INFLUENCING THE DEVELOPMENT OF THE STOCK MARKET IN THE CONDITIONS OF THE PANDEMIC OF COVID-19 VIRUS IN THE REPUBLIC OF UZBEKISTAN

	x3	x7
x3	1.0000	0.0775
x7	0.2999	1.0000

The regression matrix was obtained on the basis of a two-factor SEM model of the impact of the Covid-19 coronavirus pandemic on the development of the stock market in the Republic of Uzbekistan in Fig. 2 (Table VII).

$$Y = 264240,1 + 5,27 \cdot X_3 + 0,14 \cdot X_7 \quad (1)$$

Here:

Y - volume of daily trade turnover on the stock exchange of the Republic of Tashkent;

x3 is the total number of people who have recovered from Covid-19 coronavirus infection in the Republic of Uzbekistan.

x7 - The number of daily securities traded on the Republican Stock Exchange "Tashkent" of the Republic of Uzbekistan.

If the reliability and adequacy is checked of the equation, the Fisher criterion is F_stat = 0.006, which is less than 0.05, which means that the model is statistically significant. The p-value of each factor in the regression equation of factors x3 and x7 is 0.078 and 0.011, respectively, and at 0.1 both factors are significant.

However, the coefficient of determination is R = 0.055, which can explain only 5 percent of the total set. It can be concluded that in the context of the Covid-19 coronavirus pandemic, there are other factors that affect the daily turnover of the Republican Stock Exchange "Tashkent".

In conclusion, the increase in the number of people cured of Covid-19 coronavirus infection in the Republic of Uzbekistan by one unit will increase the daily turnover of the Republican Stock Exchange "Tashkent" by 5.27 units. An increase in the number of daily securities traded on the Republican Stock Exchange "Tashkent" of the Republic of Uzbekistan will increase the volume of daily turnover on the Republican Stock Exchange "Tashkent" by 0.005 units.

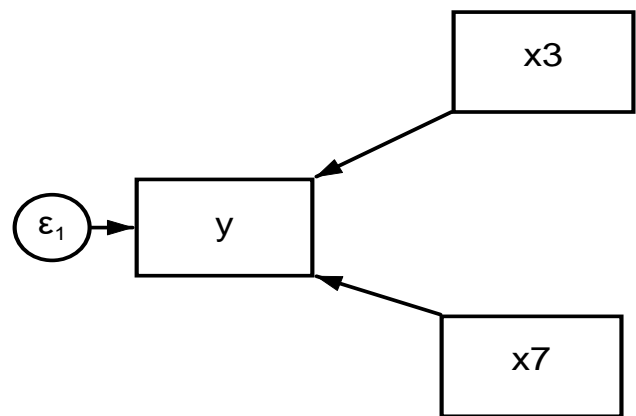


Fig. 2. Two-Factor SEM-Model of the Impact of the Covid-19 Coronavirus Pandemic on the Development of the Stock Market in the Republic of Uzbekistan.

TABLE VII. TWO-FACTOR REGRESSION MATRIX OF THE IMPACT OF THE COVID-19 CORONAVIRUS PANDEMIC ON THE DEVELOPMENT OF THE STOCK MARKET IN THE REPUBLIC OF UZBEKISTAN

Source	SS	df	MS	Number of obs = 181		
Model	1.1195e+13	178	1.7988e+12	F(8, 171) = 5.26	Prob > f = 0.1107	R-squared = 0.0718
Residual	1.8928e+14	2	1.0819e+12	Adj R-Squared = 0.0286	Root MSE = 1.0e+06	
Total	2.0047e+14	180	1.1137e+12			
<i>y</i>	Coef	Std.Err.	<i>t</i>	<i>P</i> > <i>t</i>	95% conf interval	interval
X_3	5.273639	2.970719	1.78	0.078	-0.587678	11.13496
X_7	0.0147473	0.005738	2.57	0.011	0.003423	0.026071
_cons	264240.1	107257.5	2.46	0.015	52580.26	475899.9

V. CONCLUSIONS AND SUGGESTIONS

The increase in the number of Covid-19 coronavirus infections in the Republic of Uzbekistan will lead to an increase in stock market turnover. This is because the increase in the number of people cured of the virus leads to an increase in the number of securities, resulting in an acceleration of socio-economic activity in the general population.

While Spanish researchers found that fear and uncertainty in the stock market during the Covid-19 coronavirus pandemic increased market volatility, our empirical study showed that fear and uncertainty in the Covid-19 coronavirus pandemic in developing countries weakened stock market activation.

According to the multivariate SEM model of the impact of the Covid-19 coronavirus pandemic on the development of the stock market in the Republic of Uzbekistan, the daily turnover of the country's stock exchange did not significantly affect the number of people infected, cured and died of Covid-19 coronavirus infection.

This means that the development of the Uzbek stock market is lower than in other Asian stock markets. During the Covid-19 coronavirus pandemic in the Asian, European and US stock markets, the price of shares in pharmaceuticals, IT programs and sales of companies in this field increased sharply. Therefore, it is necessary to emphasize the introduction of modern IT technologies in the stock market of Uzbekistan.

In order to accelerate the effective functioning of the stock market in Uzbekistan, it is necessary to assist in expanding the activities of investment companies, consulting firms, management companies, nominal depositors and underwriters.

ABBREVIATIONS

GST: The goods and services tax.

COVID-19: Coronavirus disease.

FTSE: The Financial Times Stock Exchange Index.

GDP: Gross domestic product.

MIB: Market index for the Borsa.

SEM-model: Structural equation Model.

ACKNOWLEDGMENTS

The views and opinions expressed in this paper are solely those of the author and are not necessarily those of the State Secretariat for Economic Affairs. We would thank Prof. Dilshodjon for his support and invaluable advice.

AUTHOR'S CONTRIBUTIONS

The author(s) read and approved the final manuscript.

FUNDING

None.

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

- [1] Arshian Sharifa, Chaker Aloui, Larisa Yarovay (2020). COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: Fresh evidence from the wavelet based approach. *International Review of Financial Analysis*, Vol.70, pp.1-9.
- [2] Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., & Viratyosinn, T. (2020). The unprecedented stock market reaction to COVID-19. Available on the link https://www.policyuncertainty.com/media/StockMarkets_COVID.pdf-24-03-2020.
- [3] Conlon, T., McGee, R. (2020) Safe haven or risky hazard? Bitcoin during the COVID-19 bear market (March 24, 2020). Available at SSRN: <https://doi.org/10.2139/ssrn.3560361>.
- [4] Corbet et al. (2020a) Corbet, S., Hou, G., Yang, H., Lucey, B. M., Les, O. (2020). Aye Corona! The contagion effects of being named corona during the COVID-19 pandemic (March 26, 2020). Available at <https://doi.org/10.2139/ssrn.3561866>.
- [5] Correia, S., Luck, S., & Verner, E. (2020). Pandemics depress the economy. Public health interventions do not: Evidence from the 1918 flu. Tech. rep. <https://doi.org/10.2139/ssrn.3561560>.
- [6] Harvey, A. C. (2020). The economic and financial implications of COVID-19 (3rd April, 2020). the Mayo Center for Asset Management at the University of Virginia Darden School of Business and the Financial Management Association International virtual seminars series. <https://www.darden.virginia.edu/mayo-center/events/virtualseminal-series>.
- [7] Ma, C., Rogers, J. H., & Zhou, S. (2020). Global economic and financial effects of 21st century pandemics and epidemics. Paper available on the link https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3565646.

- [8] Wu, K. J. Z., Xu, M., & Yang, L. (2020). Can crude oil drive the co-movement in the international stock market? Evidence from the partial wavelet coherence analysis. *The North American Journal of Economics and Finance*, 53, 101–119.
- [9] Shobhit Aggarwal, Samarpan Nawn, Amish Dugar (2020). What caused global stock market meltdown during the COVID pandemic–Lockdown stringency or investor panic? *Finance Research Letters*. November`2, pp.56-67. <https://doi.org/10.1016/j.frl.2020.101827>.
- [10] Oluwasegun B. Adekoya, Johnson A. Oliyide, Gabriel O. Oduyemi (2020). How COVID-19 upturns the hedging potentials of gold against oil and stock markets risks: Nonlinear evidences through threshold regression and markov-regime switching models. *Resources Policy*. November`9, pp.32-41. <https://doi.org/10.1016/j.resourpol.2020.101926>.
- [11] Badar Nadeem Ashraf. Stock markets' reaction to COVID-19: Cases or fatalities? // *Research in International Business and Finance* 54 (2020) 101249. (<https://doi.org/10.1016/j.ribaf.2020.101249>).
- [12] Alok Kumar Mishra, Badri Narayan Rath, and Aruna Kumar Dash. Does the Indian Financial Market Nosedive because of the COVID-19.
- [13] Outbreak, in Comparison to after Demonetisation and the GST? // *Emerging markets finance and trade*. 2020, Vol. 56, No. 10, 2162-2180. (<https://doi.org/10.1080/1540496X.2020.1785425>).
- [14] Lopatta, Kerstin and Alexander, Ernst-Kenji and Gastone, Laura Maria and Tammen, Thomas. To Report or Not to Report About Coronavirus? The Role of Periodic Reporting in Explaining Capital Market Reactions During the COVID-19 Pandemic (April 3, 2020). Available at: <http://dx.doi.org/10.2139/ssrn.3567778>.
- [15] <https://www.worldometers.info/coronavirus/> (2020).
- [16] <https://coronavirus.uz/ru> (2020).
- [17] <https://uza.uz/oz/business/dunye-davlatlari-i-tisodi-koronavirusdan-aydarazhada-zarar--27-03-2020>.
- [18] <https://nbu.uz/exchange-rates> (2020).
- [19] <https://ru.investing.com>; (2020).
- [20] <https://www.ftserussell.com> (2020).
- [21] <https://www.bloomberg.com> (2020).
- [22] <https://finance.yahoo.com> (2020).

Deep Learning based Anomaly Detection in Images: Insights, Challenges and Recommendations

Ahad Alloqmani¹, Yoosef B. Abushark², Asif Irshad Khan³, Fawaz Alsolami⁴
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Abstract—Deep learning-based anomaly detection in images has recently been considered a popular research area with numerous applications worldwide. The main aim of anomaly detection (i.e., Outlier detection), is to identify data instances that deviate considerably from the majority of data instances. This paper offers a comprehensive analysis of previous works that have been proposed in the area of anomaly detection in images through deep learning generally and in the medical field specifically. Twenty studies were reviewed, and the literature selection methodology was defined based on four phases: keyword filter, publish filter, year filter, and abstract filter. In this review, we highlight the differences among the studies included by considering the following factors: methodology, dataset, pre-processing, results and limitations. Besides, we illustrate the various challenges and potential future directions relevant to anomaly detection in images.

Keywords—Anomaly detection; outlier detection; deep learning

I. INTRODUCTION

Identifying examples that deviate from what is typical or expected is the primary goal of anomaly detection and known as outlier detection [1]. Anomaly detection in images has recently been considered a popular research area with numerous applications in different fields ranging from the video surveillance field to medical fields [2] [3]. Anomalies arise due to various reasons such as data errors or data noises but sometimes indicate a new process that was previously unseen. Thus, anomaly detection is a crucial task, especially in medical image processing.

Many researchers tended to employ deep learning to detect abnormalities in images, due to the proliferation of deep neural networks, with unprecedented results across various applications. It can also deal with complicated features such as regions of interest points by examining every pixel in an image [4] [5].

In fact, deep learning-based anomaly detection have gained prominence and have been applied to various tasks, with the help of the technologies increasingly popular in the medical sector [3] [6–9]. This is because deep learning overcomes the issue of data being imbalanced, which may result in a bias towards the majority group (i.e., the negative case). Since the medical images for the negative cases are more than the positive ones, we believe that anomaly detection can be considered a better technique to be adopted than the binary classification [9].

There are several papers from different fields in the area

of deep learning-based anomaly detection. We believe there is a gap in the literature about having reviews that state the gaps and limitations of the topic of interest of this article. Therefore, we opt to have a review article that collects and comprehensively analyzes recent works on deep learning-based anomaly detection in images. Hence, the community would be able to effortlessly understand the contributions and limitations of each study and to overcome these limitations in their future work.

This study aims to illustrate the state-of-the-art techniques for anomaly detection in images by reviewing recent studies that leverage deep learning techniques for anomaly detection. In our survey, we classify anomaly detection into two categories: general and medical fields in the context of medical anomalies. This study also discusses several factors that make the anomaly detection approach challenging. Such factors include the availability of labeled data, how to deal with noise that tends to be similar to the actual anomalies, and therefore, difficult to distinguish.

The significant contributions of this paper are as follows: (a) A comprehensive analysis of previous works that have been proposed in the area of anomaly detection in images through deep learning generally and in the medical field specifically by considering methodology, dataset, pre-processing, findings and limitations, outlining the difference between these studies. (b) Illustrate the various challenges and potential future directions relevant to anomaly detection in images.

The remainder of this article is organized as follows: the background of this study is given in Section II. In Section III we provide the necessary information for the reader to understand the rest of the article. Section IV discusses the literature selection methodology. Recent works of deep learning-based anomaly detection are reviewed in Section V. Observations and challenges are discussed in Section VI, while we conclude and provide the future work in Section VII and VIII.

II. BACKGROUND

This section explains the necessary background to understand the various elements of this article. We briefly explain the elements of the context of this review (i.e., anomaly detection, deep learning, and automated medical image diagnosis).

A. Anomaly Detection

Anomaly detection, known as outlier detection, is defined as the process of identifying data instances that deviate

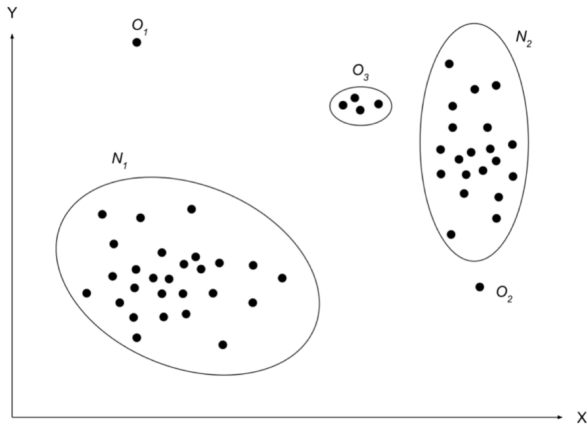


Fig. 1. Illustration of Anomalies in Two-Dimensional Dataset [5].

tremendously from other data instances [4]. As shown in Fig. 1, “N1” and “N2” are regions containing the majority of observations and are therefore considered to be normal data instance regions, while the “O3” area and the “O1” and “O2” data points are the few data points located far from the bulk of the data points. Given that “O3”, “O1”, and “O2” are therefore considered to be anomalies. They occur due to data errors but sometimes indicate a new basic process that was not previously known [5]. Anomaly detection plays an increasingly important role and is highlighted in different communities, including machine learning, computer vision, and data mining [4].

B. Deep Learning

In recent years there has been exponential development of deep learning and has been shown through several various application areas. Deep learning is considered a sub-domain of the machine learning field that aims to achieve good performance and flexibility [4]. As R. Chalapathy et al. stated in [5], deep learning achieves outstanding performance and flexibility than machine learning through learning to represent data as a nested hierarchy of concepts within the layers of a neural network. As Fig. 2 shows, deep learning outperforms the conventional approaches of machine learning considering the increased data scale [10].

C. Automated Medical Image Diagnosis

In the field of medical image processing, automated diagnosis is the primary and most important task. Automated diagnosis is based on the detection of abnormal behavior in the images [11]. Still detect abnormalities such as malignant tumors from medical images, including mammograms or CT scan, are ongoing research problems that attract a lot of attention with applications in medical diagnosis [9].

III. TERMINOLOGY

There are basic terminologies in the anomaly detection field, and they are as follows.

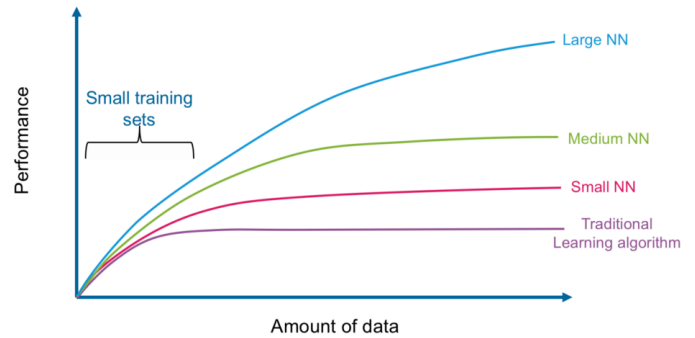


Fig. 2. Comparing the Performance of Deep Learning-based Algorithms Versus Traditional Algorithms [10].

A. Deep Learning

Deep learning is “learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower-level features” [12]. Means deep learning learns layers of features.

B. Anomaly Detection

Anomaly detection is the process of identifying data instances that deviate from what is normal or expected data [1].

C. Semi-Supervised or (one-class classification) Deep Anomaly Detection

Defined as “a technique assumes that all training instances have only one class label” [5].

D. Unsupervised Deep Anomaly Detection

Unsupervised is “a technique that used automatic labeling of unlabeled data samples” [5].

E. Normal Data

Normal data are the majority of data instances (usually be the negative data in the medical field) [5] [9].

F. Anomalous/Abnormal data

Abnormal data are the deviants in data instances (usually be the positive/diseases data in the medical field)[5] [9].

G. Anomaly Score

is “describes the level of outlierness for each data point” [5].

IV. LITERATURE SELECTION METHODOLOGY

In order to review the most important anomaly detection literature for this review, an existing selection methodology was having been adapted from [13]. This section provides a description of the process for selecting literature (see Fig. 3).

A. Keywords Filtering Stage

We started by selecting the related articles from the Google Scholar search engine, arXiv and bioRxiv using at least one of the following keywords in the title of the article: (1) anomaly detection, (2) anomaly detection in images, (3) anomaly detection in medical images, or (4) deep learning-based anomaly detection. Results from this stage 55 articles.

B. Publishers Filtering Stage

The methodology of the literature collection included article published by these publishers: (1) Springer, (2) IEEE, (3) Elsevier, (4) ACM, (5) ICLR, (6) SPIE, and (7) arXiv and bioRxiv preprints. Fig. 4 presents the percentage of articles for each publisher. Results from this stage reduced from 55 to 40.

C. Year Filtering Stage

The methodology of literature selection also focused on recent research articles in recent years by considering the following years only: (1) 2020, (2) 2019, and (3) 2018. Fig. 5 presents the percentage of articles for each year. Results from this stage reduced from 40 to 28.

D. Abstract Filtering Stage

An abstract reading was carried out in view of the 28 articles from the previous stage in order to identify only the most important articles that specifically study the deep learning-based anomaly detection in images and focus in particular on the medical field. Therefore, from the anomaly detection literature, 20 articles were chosen.

V. RECENT WORKS OF DEEP LEARNING-BASED ANOMALY DETECTION

In this paper, twenty papers on detecting anomalies in images through deep learning generally and in the medical field specifically were reviewed. Fig. 6 presents the percentage of articles for each field.

A. General Field

This section will present some previous works of anomaly detection in terms of the general field.

The authors of this research [14], proposed Deep Semi-Supervised Anomaly Detection (Deep SAD). Furthermore, they presented an information-theoretic framework for deep anomaly detection, which as minimizing the entropy of the

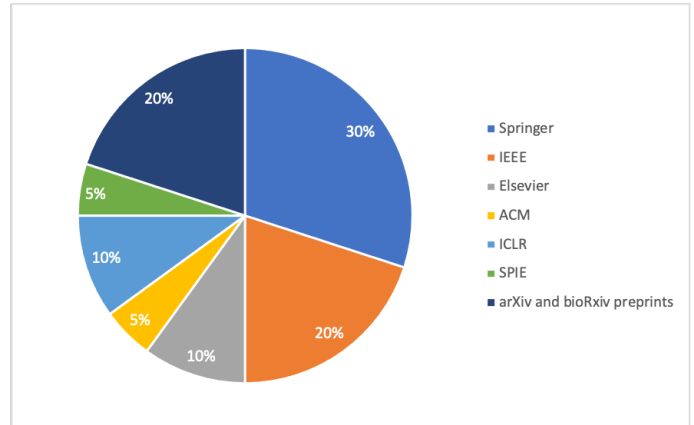


Fig. 4. Showed the Percentage Ratio of Articles for Each Publisher in 20 Articles.

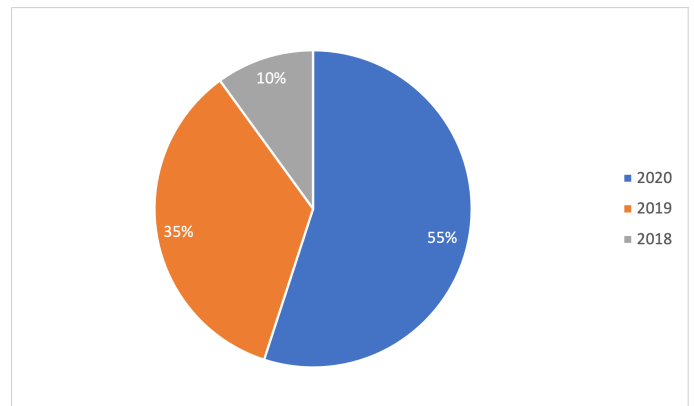


Fig. 5. Showed the Percentage Ratio of Articles for Each Year in the 20 Articles.

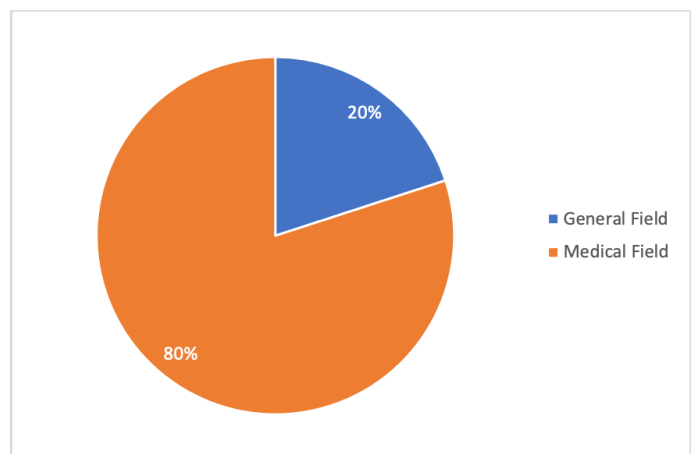


Fig. 6. Showed the Percentage Ratio of Articles for Each Field in 20 Articles.

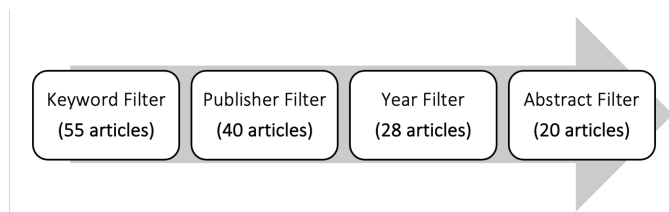


Fig. 3. Literature Selection Methodology.

latent distribution for normal data and maximizing the entropy of the latent distribution for anomalous data. The experiments were on several different public datasets and comparing their method with other previous methods. The results show that the method of this paper was on par or outperform other methods that compared it. The authors did not consider the problem of the difficulty availability of label anomalies.

This study [15] presented Iterative Training Set Refinement (ITSR), which is a novel method. An adversarial autoencoder architecture is geared to overcome the shortcomings of conventional autoencoders in the existence of anomalies in the training set. They used two public datasets, MNIST, and Fashion-MNIST datasets. The results show that their method has better accuracy than traditional autoencoders and adversarial autoencoders. However, they did not experiment with their method when there are noises in images which means do not consider preprocessing data. Also, they did not compare their result with other works or state-of-the-art methods.

This research [16] proposed a new framework and its instantiation Deviation Networks (DevNet) to take advantage of a few labeled anomalies with a prior probability to fulfill end-to-end differentiable learning of anomaly scores. Nine publicly available real data sets were used, and are from various critical fields, for example, fraud detection, disease detection, malicious URL detection, and intrusion detection. The experimental findings indicate that their current approach was more effective score than state-of-the-art competing methods. But the authors did not examine the lack of label anomalies data in the real world, particularly in medicine field.

On the contrary, using an unsupervised model is the proposed method of this paper [17], where the authors present a Deep Autoencoding Gaussian Mixture Model (DAGMM) for unsupervised anomaly detection. The experiment was applied to four public benchmark datasets and compared the results with state-of-the-art anomaly detection techniques. The results indicate that DAGMM exceeds state-of-the-art anomaly detection methods with a 14% improvement based on the standard F1 score. However, they did not test their method on images with noises to show the extent of its impact on the results.

B. Medical Field

This section will present some previous works of anomaly detection in terms of the medical field by considering the application area.

1) *Breast*: According to new research by [18], the authors introduced a new method that is a new measure for determining the effect of a particular sample on a task, allowing to detect samples outside of distribution. Their method integrated into a simple autoencoder CAE model for the abnormality recognition task. Examination of their method on Breast Magnetic Resonance Imaging (MRI) and Breast Full-Field Digital Mammography (FFDM) datasets. Experimental results demonstrate that the new method exhibits remarkable performance and outperforms the compared methods with accuracy 90.1% and 95.6% in MRI and FFDM datasets respectively. The experiments of the method are done on small datasets relatively.

The authors of this research [19] an architecture with two deep convolutional networks (R and M) proposed for

irregular tissues in mammography images. They used three public datasets, the Mammographic Image Analysis Society (MIAS) and INbreast dataset for training their method. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset to test their method. The accuracy they achieve is 76% and 86% in MIAS and INbreast datasets respectively. However, the datasets used are of small size. Moreover, they did not consider processing the whole image in one step in the model.

This study [9] designed an autoencoder based on a deep neural network to detect an anomaly in medical images based on one-class classification. The INbreast dataset is used, and the performance was 84%. Also, this paper used a small dataset. Furthermore, they did not compare their result with other works or state-of-the-art methods.

2) *Chest*: In terms of the chest area, the confidence-aware anomaly detection (CAAD) model for viral pneumonia screening from non-viral pneumonia and healthy controls have been implemented in recent research [20] into a one-class classification-based anomaly detection challenge. Their model consists of a function extractor, a module for detecting anomalies, and a module for predicting confidence. Four datasets were used, which are X-VIRAL, X-COVID, public COVID-19, and lastly combine the X-COVID and Open-COVID datasets. The results show the accuracy is 87.57%, 83.61%, 94.93%, and 84.43% for datasets respectively. The only limitation of this research is it did not try to consider comparing without data preprocessing to see if there is much difference in results or not.

This study [21] presented an abnormality detection method based on an autoencoder with uncertainty prediction. This method is able to reconstruct the image with pixel-wise uncertainty prediction. Two public chest X-ray datasets were used: RSNA Pneumonia Detection Challenge dataset and pediatric chest X-ray dataset. The area under the curve (AUC) was 89% and 78% for datasets respectively. There is no preprocessing data step.

In [22] an end-to-end architecture to determine a chest X-ray abnormal using generative adversarial one-class learning was proposed. It is similar to generative adversarial networks (GANs). Their architecture consists of a U-Net autoencoder, a CNN discriminator, and an encoder. The experiments were done on the NIH Clinical Center Chest X-ray dataset, and they achieve 80% accuracy to detect lung opacities. But their architecture results did not compare with other algorithms.

3) *Brain*: Since recently, researchers have shown an increased interest in Generative Adversarial Network (GAN) on deep learning. Accordingly, this paper [23] introduced unsupervised anomaly detection Generative Adversarial Network (MADGAN) method using multiple adjacent brain MRI slice reconstruction. This approach is capable of detecting various diseases at different stages on multi-sequence structural MRI. Two different datasets were used. The MRI dataset was extracted from the Open Access Series of Imaging Studies-3 (OASIS-3) and the second dataset was collected by the authors (National Center for Global Health and Medicine, Tokyo, Japan) which is brain metastasis and various disease MRI dataset. The results demonstrate that this method can detect anomaly detection at a very early stage with 72.7% and

at a late stage with 89.4% in terms of area under the curve (AUC). But their method results did not compare with other algorithms.

A method of using GANs trained from multi-modal magnetic resonance images (MRI) as a 3-channel input is defined and demonstrated by the authors in [24]. Their model was used to detect tumour as an anomaly. The dataset was from The Cancer Imaging Archive. The resulting accuracies that differ substantially in the size of the anomaly have been observed. The area under the receiver operator characteristic curve (AUROC) was observed to be greater than 75% for anomaly sizes greater than 4 cm². The dataset consists of 20 patients, which is very small.

In [25] proposed a semi-supervised anomaly detection model to detect brain tumor abnormalities. The model consisted of four components which are the encoder-decoder part, the discriminator, latent regularizer, and auxiliary encoder. The model first has been tested on two benchmark datasets which are MNIST and CIFAR-10 for comparison with state-of-the-art methods. Then applied the model on the HCP database and BraTS dataset. Where using normal images from the HCP database as training data and the whole BraTS 2019 dataset as the test data. The results were 93%, 79.7% for MNIST and CIFAR-10 respectively. 99.4% for the BraTS dataset. There is no preprocessing data step.

4) *Eye*: In research [26], it proposed a novel P-Net for retina image anomaly detection. Their network architecture consisted of three modules which are structure extraction from the original image module, image reconstruction module, and structure extraction from the reconstructed image module. Two datasets have been used, which are Retinal Edema Segmentation Challenge Dataset (RESC) and Fundus Multi-disease Diagnosis Dataset (iSee). The result was 92.88% and 72.45% for both datasets, respectively. There is no preprocessing data step.

This study [27] proposed a transfer-learning-based approach for unsupervised anomaly detection. The methodology used a convolutional neural network as a feature extractor and Isolation Forest anomaly detection method as a classification. Two benchmark datasets (CIFAR-10 and SVHN) were used, and two retinal fundus image datasets, which are Retinopathy of Prematurity (ROP) and Diabetic Retinopathy (DR) were used. The results were 88.2%, 55.4% for CIFAR-10 and SVHN respectively. 77% and 74.5% for the ROP and DR respectively. The authors did not try to consider comparing without data preprocessing to see if there is much difference in results or not. Furthermore, the medical imaging performance results need improvement.

5) *Abdomen*: In another research that used an unsupervised model, the authors in [28] have considered the problem of other organs than the stomach in a gastric X-ray examination, which can be noisy and cause decadence of classification performance. Therefore, they proposed a deep learning-based anomaly detection model inspired by DAGMM as an organ classification task. The experiment was on one dataset, which is gastric X-ray images, and comparing with other approaches. The results show that their model outperforms the comparison models with 95.6% in terms of sensitivity. The limitation of this paper was having a small number of stomach images with

barium leaks in the gastric X-ray examinations, which are not useful in gastritis detection.

6) *Cardiac*: Another application area of the medical field in [29] where the authors proposed the decision boundary-based anomaly detection model using improved AnoGan from ECG data. The proposed model achieves 94.75% in MIT-BIH Arrhythmia ECG dataset, which is the best performance compared with many different models. The authors did not consider testing the model without their data preprocessing to illustrate the difference ratio.

7) *Musculoskeletal*: This study [30] presented a pre-processing pipeline and survey unsupervised deep learning methods for an anomaly detection task. They were comparing these methods with each other with and without their pre-processing pipeline to demonstrate which algorithm is better for this task and also to show the effect of the presence of pre-processing pipeline on the performance. They work on a subset of the MURA dataset, which is X-Ray images of hands. The results illustrated that the best model is α -GAN based (GANs) approach with 60.7%, and the best model-based autoencoder is convolutional auto-encoder (CAE) with 57%. However, the experiments were on a small dataset because they did not use a full MURA dataset.

In [31] A new CNN model consisting of some previous CNN layers with the technique of weight standardization and a learning rate scheduler was proposed. The model name is GnCNNr, an acronym for Group Normalized Convolutional Neural Networks with Regularization. MURA dataset was used for experiments. This model was compared with the conventional deep learning methods: DenseNet, Inception, Inception v2 model. The Overall performance result was 89.9% in terms of area under the receiver operating characteristic curve (AUROC). This model was compared with only conventional deep learning methods and did not compare with other works.

The authors of this research [32] introduced a new Computer-Aided Diagnosis (CADx) model based on Deep Convolutional Neural Network (Deep CNN). This model identifies musculoskeletal abnormality detection from radiographs. Ensemble techniques were used to improve the model performance. For experiments, the MURA dataset was used with four types of study (Elbow, Finger, Humerus, and Wrist). The performance results were 86.45%, 82.13%, 87.15%, and 87.86% respectively. However, their model results did not compare with other works.

VI. DISCUSSION

A. Overview

Many studies have worked on anomaly detection algorithms. Summary of the related studies on deep learning-based anomaly detection in images is presented in Table I and II for the general and medical fields respectively. After reviewing the studies, the following was observed. First, most researchers use deep learning other than machine learning. Because deep learning has better performance and can handle the complexity of images and large datasets efficiently. Second, most researchers either in general or medical fields have used unsupervised [9] [17–20] [23, 24] [26–30], or semi-supervised [14–16] [21, 22] [25] learning methods in an anomaly detection task. Third,

TABLE I. SUMMARY OF RECENT RELATED WORKS IN THE GENERAL FIELD

[Ref.] (Year)	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[14] (2019)	Deep Semi-Supervised Anomaly Detection (Deep SAD). Feature Extraction & Classification: - (MNIST, Fashion-MNIS, CIFAR-10): convolutional neural networks (CNNs). - (benchmark datasets): Multi-Layer Perceptron (MLP) architectures.	1. MNIST: (70, 000). 2. Fashion-MNIS: (70, 000). 3. CIFAR-10: (60,000). - Other anomaly detection benchmark datasets: 4. arrhythmia: (452). 5. cardio: (1831). 6. satellite: (6435). 7. satimage-2: (5803). 8. shuttle: (49,097). 9. thyroid: (3772).	Standardize features to have zero mean and unit variance.	1. MNIST: 96.9 % 2. Fashion-MNIS: 91%. 3. CIFAR-10: 81.9%. 4. arrhythmia: 75.9%. 5. cardio: 95 %. 6. satellite: 91.5%. 7. satimage-2: 99.9%. 8. shuttle: 98.4 %. 9. thyroid: 98.6%. 1.1 MNIST-(observed anomaly type): 91% 1.2 MNIST-(unobserved anomaly type): 90% 2.1.1 Fashion -MNIST: (T-shirt vs. Boot-observed anomaly type): 90% 2.1.2 Fashion -MNIST: (T-shirt vs. Boot-unobserved anomaly type): 80% 2.2.1 Fashion -MNIST: (T-shirt vs. Pullover-observed anomaly type): 80% 2.2.2 Fashion -MNIST: (T-shirt vs. Pullover-unobserved anomaly type): 80%	The difficulty availability of label anomalies.
[15] (2019)	A novel method called Iterative Training Set Refinement (ITSR) for anomaly detection in images. Feature Extraction & Classification: Adversarial autoencoders (AAE).	1. MNIST: (70,000). 2. Fashion-MNIST: (70,000).	NA		1. No Pre-processing data. 2. There is no comparison with different algorithms.
[16] (2019)	A novel anomaly detection framework and its instantiation Deviation Networks (DevNet). Feature Extraction & Classification: Multi-Layer Perceptron (MLP) network architectures.	1. donors: (619,326). 2. census: (299,285). 3. fraud: (284,807). 4. celeba: (202,599). 5. backdoor: (95,329). 6. URL: (89,063). 7. campaign: (41, 188). 8. news20: (10,523). 9. thyroid: (7,200).	For all data sets, missing values are replaced with the mean value in the corresponding feature, and categorical features are encoded by one-hot encoding.	1. donors: 100% 2. census: 68.6% 3. fraud: 92.6% 4. celeba: 87% 5. backdoor: 96.8% 6. URL: 94.1% 7. campaign: 67.9 % 8. news20: 81.7 % 9. thyroid: 78.7 %	The difficulty availability of label anomalies.
[17] (2018)	Deep Autoencoding Gaussian Mixture Model (DAGMM). Feature Extraction & Classification: Autoencoder and Gaussian Mixture Model (GMM).	1. KDDCUP: (494,021). 2. thyroid: (3772). 3. arrhythmia: (452). 4. KDDCUP-Rev: (121,597).	One-Hot Representation to encode categorical features in (KDDCUP) dataset.	1. KDDCUP: 93.69 % 2. thyroid: 47.82% 3. arrhythmia: 49.83% 4. KDDCUP-Rev: 93.80 %	Did not comparing without data preprocessing to show the difference.

most of the researches does not leverage a limited number of labeled anomalies as prior knowledge. Therefore, using this technique in future work is a good idea to avoid identifying anomalies as data noises or uninteresting data due to the lack of prior knowledge of the anomalies of interest and to increase the model's performance, as shown in [16]. Fourth, some studies used a small dataset [9] [18, 19] [24] [30]. So, there is a lack of used large datasets in an anomaly detection task. Fifth, data preprocessing is an essential technique to obtain good performance, as shown in [30]. Some researchers considered it [18] [27–30], and others are not. Sixth and finally, most studies consider the comparison with many different algorithms to illustrate the evaluation metrics of each of them, and that is an important aspect of evaluating the effectiveness of the model.

B. Challenges

There are numerous factors that make anomaly detection very challenging. First, handling the class imbalance of normal and abnormal data. Second, availability of labeled data. Third, there is often noise in the data that appears to be close to the actual anomalies and thus difficult to differentiate them [33]. Fourth, the exact concept of the anomaly varies with different areas of application. For example, fluctuations in body temperature are a small deviation from normal and might be an anomaly in the medical field. On the other hand, fluctuations in the value of a stock with a similar deviation might be normal in the stock market domain [33]. So, it is not straightforward to adapt a method developed in one field to another.

TABLE II. SUMMARY OF RECENT RELATED WORKS IN THE MEDICAL FIELD

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre- processing	Results (%:Performance of the model used)	Limitation
[18] (2020)	Breast	A new measure for determining the effect of a particular sample on a task, allowing to detect of samples outside of distribution. Feature Extraction & Classification: Convolutional Autoencoder CAE model.	1. Breast Magnetic Resonance Imaging (MRI): (2872). 2. Breast Full-Field Digital Mammography (FFDM): (304).	Image resizing.	1. MRI: 90.1% 2. FFDM: 95.6%	Used small datasets.
[19] (2019)		An architecture with two deep convolutional networks (R and M) based adversarial training. Feature Extraction: Reconstruction Network (R): Encoder-decoder networks. Classification: Matching Network (M): involves convolution and fully connected layers.	1. Mammographic Image Analysis Society (MIAS) dataset: (322). 2. INbreast dataset: (410).	NA	1. MIAS: 76% 2. INbreast: 86%	1. Used small datasets. 2. No Preprocessing data. 3. No process the whole image in one step.
[9] (2018)		An autoencoder based on a deep neural network. Feature Extraction & Classification: Autoencoder model.	INbreast dataset: (410).	NA	84%	1. Used small datasets. 2. No Preprocessing data. 3. There is no comparison with different algorithms.
[20] (2020)	Chest	Confidence-aware anomaly detection (CAAD) Feature Extraction: EfficientNet. Classification: Multi-Layer Perceptron (MLP) network architecture for anomaly detection network and four layers for Confidence prediction network.	1. X-VIRAL: (43,370). 2. X- COVID: (213). 3. Public COVID-19: (519). 4. Combine the X- COVID and Open-COVID: (2,706).	1. Image resizing. 2. Augmentation.	1. X-VIRAL: 87.57% 2. X- COVID: 83.61% 3. Public COVID-19: 94.93% 4. X- COVID and Open-COVID: 84.43%	Did not comparing without data preprocessing to show the difference.
[21] (2020)		Autoencoder with pixel-wise uncertainty prediction. Feature Extraction & Classification: Autoencoder.	1. RSNA Pneumonia Detection Challenge dataset: (26,684). 2. Pediatric chest X-ray dataset: (5,856).	NA	1. RSNA: 1.1 (normal vs. lung opacity): 89% 1.2 (normal vs. not normal): 78% 1.3 (normal vs. all) - (lung opacity and not normal): 83% 2. Pediatric: 78%	No Preprocessing data
[22] (2019)		An end-to-end architecture to determine a chest X-ray abnormal using generative adversarial one-class learning. Feature Extraction & Classification: U-Net autoencoder, CNN discriminator and second encoder.	The NIH Clinical Center Chest X-ray dataset: (112,120).	Image resizing.	1. Normal vs. Abnormal: 84.1% 2. Normal vs. Ling opacities: 80.2%	There is no comparison with different algorithms.

VII. CONCLUSION

This article presents a systematic study of recent research in general and medical fields on anomaly detection in images by considering methodology, dataset, pre-processing, findings and limitations, outlining the difference between these studies. The majority of anomaly detection studies focus on the medical field since it is the best technique than binary classification to cope with imbalanced data that is an issue in medical applications. The study concludes that most researchers used unsupervised or semi-supervised for anomaly detection. Fur-

ther, most researchers used deep learning other than machine learning; Deep learning has better performance and can efficiently handle the complexity of images and large datasets. The limitation of this research is the limit of the number of literatures researched. While the authors used several databases, the ones used in the extensive index might not be exhaustive ones.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre- processing	Results (%:Performance of the model used)	Limitation
[23] (2020)	Brain	Unsupervised Medical Anomaly Detection GAN using multiple adjacent brain MRI slice reconstruction (MADGAN). Feature Extraction & Classification: GAN with include U-Net.	1. MRI dataset extracted from the Open Access Series of Imaging Studies-3 (OASIS-3): (1,606 scans). 2. Brain metastasis and various disease MRI dataset collected by the authors (National Center for Global Health and Medicine, Tokyo, Japan): (193 scans).	NA	At a very early stage: 72.7% At a late stage: 89.4%	There is no comparison with different algorithms.
[24] (2020)		A method of using GANs trained from multi-modal magnetic resonance images (MRI) as a 3-channel input. Feature Extraction & Classification: GAN	Multi-modal magnetic resonance brain images MRI dataset from The Cancer Imaging Archive: (308).	NA	(AUC) was observed to be greater than 75% for anomaly sizes greater than 4 cm ² . Sensitivity (Sen): All tumours: 99% Area >4 cm ² : 99% Area >7 cm ² : 97%	1. Used small datasets. 2. There is no comparison with different algorithms.
[25] (2020)		A semi-supervised anomaly detection model to detect brain tumor abnormalities. Feature Extraction & Classification: The GAN-style architecture: the encoder-decoder part, the discriminator, auxiliary encoder, and latent regularizer.	1. MNIST dataset: (70,000). 2. CIFAR-10 dataset: (60,000). 3. HCP database -Training only-: (65 healthy patients). 4. BraTS dataset: (335 patients).	NA	1. MNIST: 93% 2. CIFAR-10: 79.7% 3. BraTS: 99.4%	No Preprocessing data.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre- processing	Results (%:Performance of the model used)	Limitation
[26] (2020)	Eye	A novel P-Net methodology is proposed by the researcher for the detection of anomalies in retina images. Feature Extraction & Classification: U-Net autoencoder and Discriminator architecture.	1. Retinal Edema Segmentation Challenge Dataset (RESC): NA. 2. Fundus Multi-disease Diagnosis Dataset (iSee): (10,000).	NA	1. RESC: 92.88% 2. iSee: 72.45%	No Preprocessing data.
[27] (2019)		This research applied transfer-learning based method for unsupervised anomaly detection. Feature Extraction: CNN: Inception-ResNet-v2 network Classification: Isolation unsupervised anomaly detection method (Isolation Forest method).	1. CIFAR-10 dataset: (60,000). 2. SVHN dataset (99,289): 3. Retinopathy of Prematurity (ROP): (5511). 4. Diabetic Retinopathy (DR): (11,741).	CIFAR-10 & SVHN: Rescaling the images to [0, 1]. ROP & DR: Squared cropped to cut the neutral background and resized images to 256 pixels.	1. CIFAR-10: 88.2% 2. SVHN: 55.4% 3. ROP: 77% 4. DR: 74.5%	1. Did not comparing without data preprocessing to show the difference. 2. Medical imaging performance results need improvement.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[28] (2020)	Abdomen	Deep Autoencoding Gaussian Mixture Model (DAGMM). Feature Extraction & Classification: Convolutional Autoencoder and Gaussian Mixture Model (GMM).	Gastric X-ray dataset: (48,012).	Image resizing.	- Sensitivity (Sen): 95.6 % - Specificity (Spe): 98% - Harmonic mean of sensitivity and specificity (HM): 96.8%	There are still a minimal amount of stomach images in the gastric X-ray examinations with barium leakage that are not successful in addressing gastritis detection.
[29] (2020)	Cardiac	This research suggested a decision boundary-based Anomaly Detection model using improved AnoGan that uses ECG dataset. Feature Extraction & Classification: AnoGan.	MIT-BIH Arrhythmia ECG dataset: (85,717).	1. Filtering using Hamilton algorithm. 2. R-peak Detection and ECG Data Segmentation for signal processing. 3. Gray Scale Conversion and resize for image processing. 4. Segmentation is performed in the range between 0.3 seconds and 0.4 seconds on the basis of R-peak.	94.75%	Did not comparing without data preprocessing to show the difference.

TABLE II. CONTINUED

[Ref.] (Year)	Application Area	Methodology	Dataset (#: Sample size)	Pre-processing	Results (%:Performance of the model used)	Limitation
[30] (2020)	Musculoskeletal	Unsupervised anomaly detection in X-ray images. Feature Extraction & Classification: CAE, VAE, DCGAN, BiGAN, α -GAN.	MURA dataset containing only X-ray images of hands: (5,543).	Introduced preprocessing pipeline 1. Cropping. 2. Localization (single shot multibox detector - SSD) 3. Hand Segmentation using Photoshop's. 4. Augmentation. 5. Padding & Centring. 6. Min-Max Normalization	1. CAE: 57 % 2. VAE: 48.3% 3. DCGAN: 53% 4. BiGAN: 54.9% 5. α -GAN: 60.7%	1. The segmentation manually. 2. Used small datasets.
[31] (2020)		A Group Normalized Convolutional Neural Networks with Regularization (GnCNNr) model. Feature Extraction & Classification: New CNN model- (GnCNNr).	MURA dataset containing images of hand, wrist, humerus, shoulder, elbow, finger and forearm: (40,561).	1. Images used are of fixed size. 2. Increased channels. 3. Normalization. 4. Data Augmentation.	1. Hand: 83.5% 2. Wrist: 93.2% 3. Humerus: 92.4% 4. Shoulder: 85.6% 5. Elbow: 90.6% 6. Finger: 88.8% 7. Forearm: 92.6%	There is no comparison with different works, just comparing with conventional deep learning methods.
[32] (2019)		This research proposed a novel Computer-Aided Diagnosis (CADx) model based on Deep Convolutional Neural Network (Deep CNN). Feature Extraction & Classification: VGG-19 and ResNet.	MURA dataset containing mages of elbow, finger, humerus, and wrist: (22,938).	1. Image normalization. 2. Gaussian blur. 3. Histogram equalization. 4. Adaptive thresholding.	1. Elbow: 86.45% 2. Finger: 82.13% 3. Humerus: 87.15% 4. Wrist: 87.86%	There is no comparison with different works.

VIII. FUTURE WORK

As future work, we would establish an anomaly detecting mechanism utilizing deep learning techniques for detecting breast cancer.

REFERENCES

- [1] N. Sarafijanovic-Djukic and J. Davis, "Fast distance-based anomaly detection in images using an inception-like autoencoder," in *International Conference on Discovery Science*, pp. 493–508, Springer, 2019.
- [2] X. Xie, C. Wang, S. Chen, G. Shi, and Z. Zhao, "Real-time illegal parking detection system based on deep learning," in *Proceedings of the 2017 International Conference on Deep Learning Technologies*, pp. 23–27, 2017.
- [3] W. Shi, G. Yan, Y. Li, H. Li, T. Liu, C. Sun, G. Wang, Y. Zhang, Y. Zou, and D. Wu, "Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty," *NeuroImage*, vol. 223, p. 117316, 2020.
- [4] G. Pang, C. Shen, L. Cao, and A. v. d. Hengel, "Deep learning for anomaly detection: A review," *arXiv preprint arXiv:2007.02500*, 2020.
- [5] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.
- [6] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimesch, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.
- [7] H. Choi, S. Ha, H. Kang, H. Lee, D. S. Lee, A. D. N. Initiative, *et al.*, "Deep learning only by normal brain pet identify unheralded brain anomalies," *EBioMedicine*, vol. 43, pp. 447–453, 2019.
- [8] S. Xu, H. Wu, and R. Bie, "Cxnet-m1: Anomaly detection on chest x-rays with image-based deep learning," *IEEE Access*, vol. 7, pp. 4466–4477, 2018.
- [9] Q. Wei, Y. Ren, R. Hou, B. Shi, J. Y. Lo, and L. Carin, "Anomaly detection for medical images based on a one-class classification," in *Medical Imaging 2018: Computer-Aided Diagnosis*, vol. 10575, p. 105751M, International Society for Optics and Photonics, 2018.
- [10] A. Ng, "Nuts and bolts of building ai applications using deep learning," *NIPS Keynote Talk*, 2016.
- [11] M. Abbass, K.-C. Kwon, N. Kim, S. A. Abdelwahab, N. Haggag, F. Ibrahim, Y. Mahrous, A. Seddik, A. Khalil, Z. Elsherbeeney, *et al.*, "Anomaly detection from medical signals and images using advanced convolutional neural network," *Research Square*, 2020.
- [12] Y. Bengio, *Learning deep architectures for AI*. Now Publishers Inc, 2009.
- [13] D. A. Alboaneen, D. Alsaffar, A. Alateeq, A. Alqahtani, A. Alfahhad, B. Alqahtani, R. Alamri, and L. Alamri, "Internet of things based smart mirrors: A literature review," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–6, IEEE, 2020.
- [14] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.
- [15] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust anomaly detection in images using adversarial autoencoders," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 206–222, Springer, 2019.
- [16] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 353–362, 2019.
- [17] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International Conference on Learning Representations*, 2018.
- [18] L. Gao and S. Wu, "Response score of deep learning for out-of-distribution sample detection of medical images," *Journal of Biomedical Informatics*, vol. 107, p. 103442, 2020.
- [19] M. Ahmadi, M. Sabokrou, M. Fathy, R. Berangi, and E. Adeli, "Generative adversarial irregularity detection in mammography images," in *International Workshop on Predictive Intelligence In MEDicine*, pp. 94–104, Springer, 2019.
- [20] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, *et al.*, "Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection," *IEEE transactions on medical imaging*, 2020.
- [21] Y. Mao, F.-F. Xue, R. Wang, J. Zhang, W.-S. Zheng, and H. Liu, "Abnormality detection in chest x-ray images using uncertainty prediction autoencoders," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 529–538, Springer, 2020.
- [22] Y.-X. Tang, Y.-B. Tang, M. Han, J. Xiao, and R. M. Summers, "Abnormal chest x-ray identification with generative adversarial one-class classifier," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1358–1361, IEEE, 2019.
- [23] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z. A. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, "Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction," *arXiv preprint arXiv:2007.13559*, 2020.
- [24] S. Benson and R. Beets-Tan, "Gan-based anomaly detection in multi-modal mri images," *bioRxiv*, 2020.
- [25] N. Wang, C. Chen, Y. Xie, and L. Ma, "Brain tumor anomaly detection via latent regularized adversarial network," *arXiv preprint arXiv:2007.04734*, 2020.
- [26] W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," In: *Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, vol. 12365, 2020.
- [27] K. Ouardini, H. Yang, B. Unnikrishnan, M. Romain, C. Garcin, H. Zenati, J. P. Campbell, M. F. Chiang, J. Kalpathy-Cramer, V. Chandrasekhar, *et al.*, "Towards practical unsupervised anomaly detection on retinal images," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pp. 225–234, Springer, 2019.

- [28] R. Togo, H. Watanabe, T. Ogawa, and M. Haseyama, "Deep convolutional neural network-based anomaly detection for organ classification in gastric x-ray examination," *Computers in Biology and Medicine*, vol. 123, p. 103903, 2020.
- [29] D.-H. Shin, R. C. Park, and K. Chung, "Decision boundary-based anomaly detection model using improved anogan from ecg data," *IEEE Access*, vol. 8, pp. 108664–108674, 2020.
- [30] D. Davletshina, V. Melnychuk, V. Tran, H. Singla, M. Berrendorf, E. Faerman, M. Fromm, and M. Schubert, "Unsupervised anomaly detection for x-ray images," *arXiv preprint arXiv:2001.10883*, 2020.
- [31] M. Goyal, R. Malik, D. Kumar, S. Rathore, and R. Arora, "Musculoskeletal abnormality detection in medical imaging using gncnnr (group normalized convolutional neural networks with regularization)," *SN Computer Science*, vol. 1, no. 6, pp. 1–12, 2020.
- [32] T. C. Mondol, H. Iqbal, and M. Hashem, "Deep cnn-based ensemble cadx model for musculoskeletal abnormality detection from radiographs," in *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 392–397, IEEE, 2019.
- [33] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

The Evaluation of User Experience Testing for Retrieval-based Model and Deep Learning Conversational Agent

Pui Huang Leong¹, Ong Sing Goh², Yogan Jaya Kumar³, Yet Huat Sam⁴, Cheng Weng Fong⁵

Faculty of Computing and Information Technology, Tunku Abdul Rahman University College (TAR UC), Johor, Malaysia^{1,5}
Centre for Advanced Computing Technology, Faculty of Information and Communication Technology^{2,3}

Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia^{2,3}

Faculty of Accountancy, Finance and Business, Tunku Abdul Rahman University College (TAR UC), Johor, Malaysia⁴

Abstract—The use of a conversational agent to relay information on behalf of individuals has gained worldwide acceptance. The conversational agent in this study was developed using Retrieval-based Model and Deep Learning to enhance the user experience. Nevertheless, the successfulness of the conversational agent could only be determined upon the evaluation. Thus, the testing was performed in the quantitative approach via questionnaire survey to capture user experience upon the usage of the conversational agent in terms of Usability, Usefulness and Satisfaction. The questionnaire survey was tested via statistical tool for reliability and validation test and proven to be carried out. The test results indicate positive experience towards the usage of the conversational agent and the outcome of the testing showed promising results and proof the success of this study, with immense contributions to the field of conversational agent.

Keywords—Conversational agent; retrieval-based model; deep learning; user experience testing; usability; usefulness; satisfaction

I. INTRODUCTION

User experience may be unique to the extent that it affects human interpretation and feeling regarding a service or program. In brief, user experience was about how a consumer communicates with a device and its interactions. This broadness means that the user experience included several facets, when documenting user interaction while utilising a conversational agent. User Experience testing was conducted to capture user experience upon interacting with the conversation agent. By the end of the testing, questionnaire survey was handed out to obtain the feedback and satisfaction towards the usage of the conversational agent.

The international agreement on the ergonomics of contact between human beings as stated in the ISO 9241-210:2010 depicted the User Experience as the expectations and reactions of an individual arising from the usage or expectation of service, device or system. User experience involved the feelings, desires, attitudes, physical and psychological reactions, habits of all users that emerge before, during and after the usage. According to [1], when measuring user experience while using a chatbot, user experience could be separated into three specific needs, namely Usefulness, Usability and Satisfaction.

User experience testing was a process in which the interface and the chatbot features were verified by end users who execute specific tasks under practical environments. This test aimed to assess the user experience in terms of Usability, Usefulness and Satisfaction of the conversational agent and to ascertain if the application was functional. At the end of the testing, a survey was carried out via Google Form to gather user' response and satisfaction towards the usage of the conversational agent. Prior to the released of the questionnaire survey for User Experience Testing, validation of the research instruments will be conducted to ensure the survey was ready to be escalated for the real testing.

The next section discussed the three aspects of the User Experience Testing in details. Section III discussed the research instruments questionnaires followed by Section IV to discuss the validations of the research instruments. Next, sample size population was discussed to provide insight on how the total number of respondents were determined. Section VI discussed the testing and analysis followed by the last section to discuss the conclusion of the study.

II. USER EXPERIENCE TESTING

Generally, the user experience was measured in three aspects, namely Usefulness, Usability and Satisfaction via the quantitative method in this research. The user experience testing was scoped into three measures as illustrated in Fig. 1.

The following section discussed the Usability measure, Usefulness measure and Satisfaction measure of the User Experience Testing in depth.



Fig. 1. User Experience (UX) testing [1].

A. Usability Measure

Usability was part of the broader phrase "user experience" which refers to a software or service was readily viewed or utilised. The international standard ISO 9241-11:2018 concept of usability was: "the degree to which a service or product may be used by specified users to accomplish defined goals with effectiveness, efficiency and satisfaction in a defined context of use." The use of structured surveys was a common and cost-effective method for usability assessments. A typical usability assessment known as Usability Metric for User Experience (UMUX) as shown in Fig. 2 has been used to assess usability.

According to [2], the Usability Metric for User experience was versatile for a larger user experience variable to function as a usability element. UMUX was used in this study to evaluate the usability of user experience. The authors in [5] and [6] have been adopting UMUX to measure the usability experience of the users.

B. Usefulness Measure

Usefulness was described as being useful when it comes to quality or fact. In Technology Acceptance Model (TAM), perceived usefulness has been identified as one of the variables influencing the usage and adoption by specific users of information systems and technologies. TAM, founded by [3], was one of the most common methods of analysis to forecast the usage and recognition by specific consumers of information systems and technologies.

TAM as in Fig. 3 has been extensively examined and validated by numerous experiments that investigate the individual behaviors in acceptance of technology in diverse structures of information systems. TAM Model described there were two measures which were essential in the study of computer use behaviors, namely perceived usefulness and perceived ease of use. The author in [3] described perceived usefulness as the subjective likelihood of the prospective customer that utilising a particular application program would increase the efficiency of his or her work or existence. Perceive ease of use could be described as the degree to which the prospective consumer considers the target program to be effortless.

1.	[This system's] capabilities meet my requirements.	1	2	3	4	5	6	7	Strongly Disagree	Strongly Agree
2.	Using [this system] is a frustrating experience.	1	2	3	4	5	6	7	Strongly Disagree	Strongly Agree
3.	[This system] is easy to use.	1	2	3	4	5	6	7	Strongly Disagree	Strongly Agree
4.	I have to spend too much time correcting things with [this system].	1	2	3	4	5	6	7	Strongly Disagree	Strongly Agree

Fig. 2. Usability Metric for user Experience [2].

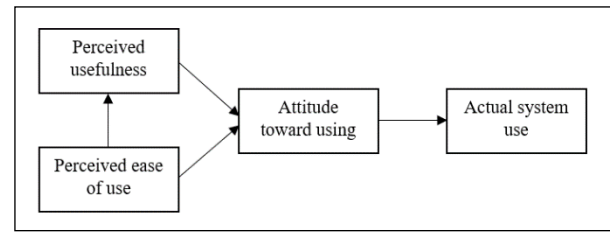


Fig. 3. Technology Acceptance Model [3].

Ease of use and perceived usefulness, according to TAM, were the most significant determinants of the actual use of the system. According to [7], TAM was amongst the essential individualistic analytical methods related to the implementation of information and communication technology (ICT). Moreover, researchers in [8], [9], [10] and [11] used TAM to test their acceptance of technologies through perceived usefulness and perceived ease of use. The relative utility of TAM was in line with the essence of this study and would be used as one of the assessing aspects of user experience.

C. Satisfaction Measure

According to the Cambridge English Dictionary, satisfaction was defined as the pleasant feeling when an individual receives something he or she wanted, or when the individual has done something he or she wanted to do. Satisfaction was one of the crucial aspects to measure user experience. The authors in [4] developed an integrated open-domain question-answering framework to test the dialogue layout that involved evaluating one of the primary factors in assessing user experience, which was the user's overall satisfaction. Two questions adapted from [4] were used to capture the overall satisfaction of the user. The next section discussed the research instruments questionnaires.

III. RESEARCH INSTRUMENTS QUESTIONNAIRES

Generally, the UMUX survey structure was used to measure the usability aspect. The UMUX questionnaire's benefit was that it comprised of only four reliable and relevant questions [2]. Moreover, the usefulness questions would be derived from the TAM questionnaire. The TAM questionnaire has the benefits of being a structured and widely employed test for measuring usefulness and ease of use [3]. This research adopting the questions to access usefulness. Finally, the questions to measure satisfaction aspect was retrieved from [4]. Table I illustrated the summary of the User Experience Testing questionnaires in this study.

There was a total of 16 questions in the questionnaire. The first part of the questions was asking user pertaining to the demographic to capture the age, education, occupation as well as if the user has ever used a chatbot. The next twelve questions derived from the questions from the three aspects to measure user experience upon the usage of the chatbot, namely the Usability, Usefulness and Satisfaction. These questions measured overall user experience upon the usage of the chatbot. All these questions were closed-ended type, and the data measure type was quantitative as the questions were prompted via the Likert-scale, ranging from strongly disagree (score-1) to strongly agree (score-5) which could be quantified

using numbers. The next section discussed the validation of the research instruments prior to the release of the User Experience Testing. Table I showed the summary of the User Experience Testing Questionnaires.

TABLE I. USER EXPERIENCE TESTING QUESTIONNAIRES

Measure	Questions	Source
Demographic	<ul style="list-style-type: none"> What was your age? What was your highest education? What was your occupation? Have you ever used a chatbot before? 	[1]
Usability	<ul style="list-style-type: none"> The chatbot's capability meet my requirements. Using this chatbot was a frustrating experience. This chatbot was easy to use. I have to spend too much time correcting things with this chatbot. 	[2]
Usefulness	<ul style="list-style-type: none"> Because of this chatbot, I could quickly execute the task (retrieve answer). This chatbot makes it hard to execute the task (retrieve answer). Because of this chatbot, I could effectively execute the task (retrieve answer). This chatbot was useless. 	[1], [3]
Satisfaction	<ul style="list-style-type: none"> Did you get all the information you wanted using the chatbot? Do you think the chatbot understood what you asked? Overall, were you satisfied with the chatbot? Do you think you would use this chatbot again? 	[4]

The next section discussed the validation of the research instruments to determine the reliability and validity of the questionnaire prior to the release of the actual survey to the respondents.

IV. VALIDATION OF RESEARCH INSTRUMENTS

User experience testing was a process in which the interface and the chatbot features were verified by end users who execute specific tasks under practical environments. This test aimed to assess the user experience in terms of Usability, Usefulness and Satisfaction of the conversational agent and to ascertain if the application was functional. At the end of the testing, a survey was carried out via Google Form to gather user' response and satisfaction towards the usage of the conversational agent.

In addition to this, a pilot test has been carried out to ascertain the validity and reliability of the questionnaire survey whereby 60 respondents have been selected. According to [12], the number of respondents in the pilot test was determined by the total number of variables tested in the questionnaire. There were three items to be tested in the questionnaire which were Usability, Usefulness and Satisfaction upon the usage of the conversational agent. The following Table II summarized the descriptive analysis of the pilot test.

TABLE II. SUMMARY OF DESCRIPTIVE ANALYSIS

Descriptive Analysis	Respondent	Total
Age		
• 18 to 24	59 (98.3%)	60
• 25 to 34	1 (1.7%)	
Education Level		
• SPM	60 (100%)	60
Occupation		
• Student	60 (100%)	60
Used Conversational Agent before		
• Yes	59 (98.3%)	60
• No	1 (1.7%)	

According to Table II, 60 respondents were students. 59 of the respondents were aged between 18 to 24 and 1 respondent was aged 25 to 34 years old. The highest educational level of the respondents was SPM and 59 respondents have experienced using conversational agent before. Furthermore, the reliability test and the validity test has been conducted and reported in Table III.

TABLE III. SUMMARY OF RELIABILITY TEST AND VALIDITY TEST

Measure	Cronbach's Alpha	KMO and Bartlett's Test
Usability	0.787	0.638**
Usefulness	0.780	0.559**
Satisfaction	0.793	0.663**
All items	0.895	0.720**

(** indicates the test was significant at 0.01 level)

The reliability test for each item has shown that the Cronbach's Alpha value for each item were more than 0.6. Furthermore, as for the KMO and Bartlett's test for validity test, all items in the questionnaire has achieved 0.720 and it was statistically significant at 0.01 level. Therefore, based on the reliability test and validity test, the questionnaire survey was suitable to be progressed to the actual survey.

V. SAMPLE SIZE POPULATION

The sample size population of respondents used in past studies pertaining to chatbot research was explored. In the study conducted by [13], a total of 169 users were participated in the study to investigate the impact of introducing language style to e-commerce chatbots to improve customer satisfaction, determined customer interest in the item and determined user interaction with the service provided by the conversational agent. Apart of this, a total of 105 participants engaged in a survey performed by [14] to test the chatbot customer service for the Venice Airport with the specially crafted modular system. A group of 101 undergraduates engaged in the study carried out by [15] to evaluate if the proposed novel paradigm enabled the users to nurture companion chatbot via developmental of artificial intelligence techniques.

Moreover, a total of 161 Korean students from major metropolitan universities in Korea engaged in research undertaken by [16] to indicate if the Chatbot e-service managed to provide interactive and engaging customer service encounters. Besides, in the test conducted by [17], 100 respondents were randomly chosen to signify the suggested

system based on certain abstract concepts, which could be applied to satisfy the necessary capabilities of the industry. In comparison, a total of 96 undergraduate computer science students engaged in research undertaken by [18] to recognise conversational agents for academically successful interactions, allowing learners to sustain effective peer dialogue in a range of learning environments.

The abovementioned evaluation on conversational agents indicated that the testing was carried out using non-probability sampling in which the approximate respondents ranging from 96 to 169 were used to carry out the testing. There was no clear generic measurement of the total number of respondents used. The sample size of this study was calculated based on the sample size formula [12], [19], refer to (1) to resolve this concern.

$$n = \frac{\left[\frac{z^2 * p(1-p)}{e^2} \right]}{\left[1 + \frac{z^2 * p(1-p)}{e^2 N} \right]} \quad (1)$$

Based on the sample size formula, n refers to the sample size, z denotes the z -score of confidence level, N denotes the population size, e denotes the margin of error, and p denotes the standard deviation. The confidence level is set at 95% with the z -score of 1.96. According to the estimation of the sample size from (1), a total of 300 users was selected to carry out the user experience testing.

VI. TESTING AND ANALYSIS

The Demographic test results were reported in Table IV. The complete graph for the demographic was then explained further in this section. Based on the survey, 83% of the respondents aged between 18 to 24 are students with SPM as the highest education which constituent to 82.7%. Moreover, 97.7% of the respondents have used chatbot before. Next, Fig. 4 showed the summary of User Experience Testing captured for Usability, Usefulness and Satisfaction. There were total of four questions for each of the user experience parameters with the mixture of positive-typed questions and negative-typed questions to prevent random answer selection by users. The survey was capture via Likert-scale ranging from score-1 to score-5 to determine the average and standard deviation of the user experience.

In order to capture reliable data, the questions were formed with the mixture of positive-type-question and negative-type-question to prevent random answer selection by users. There were total of eight positive-type-question and four negative-type-question. The positive or negative indicators could be seen next to the question number in Fig. 4. The data was quantified via the Likert-scale, ranging from strongly disagree (score-1) to strongly agree (score-5). Table V showed the total average score for positive-type-question and negative-type-question.

The results in Table V and Fig. 5 depicted that question 1, question 3, question 5, question 7, question 9, question 10, question 11 and question 12 managed to achieve the total average score of 4.74 over 5. As these questions were positive-type-question and the average was achieved more than 4.7 which was towards the strongly agree score-5 in Likert-scale,

this indicated users show positive experience towards the usage of the conversational agent. On the other hand, question 2, question 4, question 6 and question 8 were negative-type-question and each question managed to achieve the total average score of 1.24 over 5. Contrary to a positive-type-question, the lower number for negative-type-question in Likert-scale showed that users somehow deny the usage of the communication agent constitutes poor experience. Consequently, the findings of the User Experience Testing indicated that users were having positive experience of utilizing the conversational agent.

TABLE IV. SUMMARY OF DEMOGRAPHIC TEST RESULTS

Question	Options	Percentage (%)	Number of respondents
What was your age?	Below 18	0.0%	0
	18 to 24	83.0%	249
	25 to 34	7.0%	21
	35 to 44	9.3%	28
	45 to 54	0.7%	2
	55 above	0.0%	0
What was your highest education?	Diploma	1.0%	3
	Degree	8.0%	24
	Master	6.7%	20
	PhD	1.0%	3
	STPM	0.0%	0
	SPM	82.7%	248
	Other	0.6%	2
What was your occupation?	Academician	8.0%	24
	Administrator	8.7%	26
	Programmer	0.0%	0
	Engineer	0.0%	0
	Designer	0.0%	0
	Salesperson	0.0%	0
	Businessman	0.0%	0
	Student	83.3%	250
	Other	0.0%	0
Have you ever used a chatbot before?	Yes	97.7%	293
	No	2.3%	7

TABLE V. SUMMARY OF USER EXPERIENCE TESTING FOR POSITIVE-TYPE-QUESTION AND NEGATIVE-TYPE-QUESTION

Positive-type-question (+)		Negative-type-question (-)	
Question	Average score	Question	Average score
Q1	4.74	Q2	1.24
Q3	4.75	Q4	1.25
Q5	4.75	Q6	1.24
Q7	4.75	Q8	1.22
Q9	4.73		
Q10	4.71		
Q11	4.73		
Q12	4.73		
Total average	4.74/5	Total average	1.24/5

User experience parameters	No.	Questions	Type (Positive/Negative)	Average	Standard deviation	Percentage (number of respondents)				
						1 Strongly disagree	2 Disagree	3 Neutral	4 Agree	5 Strongly agree
Usability	Q1.	The chatbot's capabilities meet my requirements.	(+)	4.74	0.77	2.7% (8)	0.7% (2)	1.7% (5)	9.7% (29)	85.3% (256)
	Q2.	Using this chatbot was a frustrating experience.	(-)	1.24	0.80	89.7% (269)	4.0% (12)	2.0% (6)	1.7% (5)	2.7% (8)
	Q3.	This chatbot was easy to use.	(+)	4.75	0.74	2.0% (6)	1.3% (4)	1.7% (5)	9.7% (29)	85.3% (256)
	Q4.	I have to spend too much time correcting things with this chatbot.	(-)	1.25	0.80	88.7% (266)	5.0% (15)	2.0% (6)	1.7% (5)	2.7% (8)
Usefulness	Q5.	Because of this chatbot, I could quickly execute the task (retrieve answer).	(+)	4.75	0.74	2.0% (6)	1.3% (4)	1.7% (5)	9.7% (29)	85.3% (256)
	Q6.	This chatbot makes it hard to execute the task (retrieve answer).	(-)	1.24	0.80	89.7% (269)	4.0% (12)	2.0% (6)	1.7% (5)	2.7% (8)
	Q7.	Because of this chatbot, I could effectively execute the task (retrieve answer).	(+)	4.75	0.75	2.3% (7)	1.0% (3)	1.7% (5)	9.3% (28)	85.7% (257)
	Q8.	This chatbot was useless.	(-)	1.22	0.78	91% (273)	2.7% (8)	2.0% (6)	2.0% (6)	2.3% (7)
Satisfaction	Q9.	Did you get all the information you wanted using the chatbot?	(+)	4.73	0.76	2.3% (7)	1.0% (3)	1.7% (5)	11.3% (34)	83.7% (251)
	Q10.	Do you think the chatbot understood what you asked?	(+)	4.71	0.80	2.7% (8)	1.3% (4)	1.7% (5)	10.7% (32)	83.7% (251)
	Q11.	Overall, were you satisfied with the chatbot?	(+)	4.73	0.77	2.7% (8)	0.7% (2)	1.7% (5)	10.7% (32)	84.3% (253)
	Q12.	Do you think you would use this chatbot again?	(+)	4.73	0.83	3.3% (10)	0.7% (2)	2.0% (6)	8% (24)	86% (258)

Fig. 4. Summary of user Experience Testing.

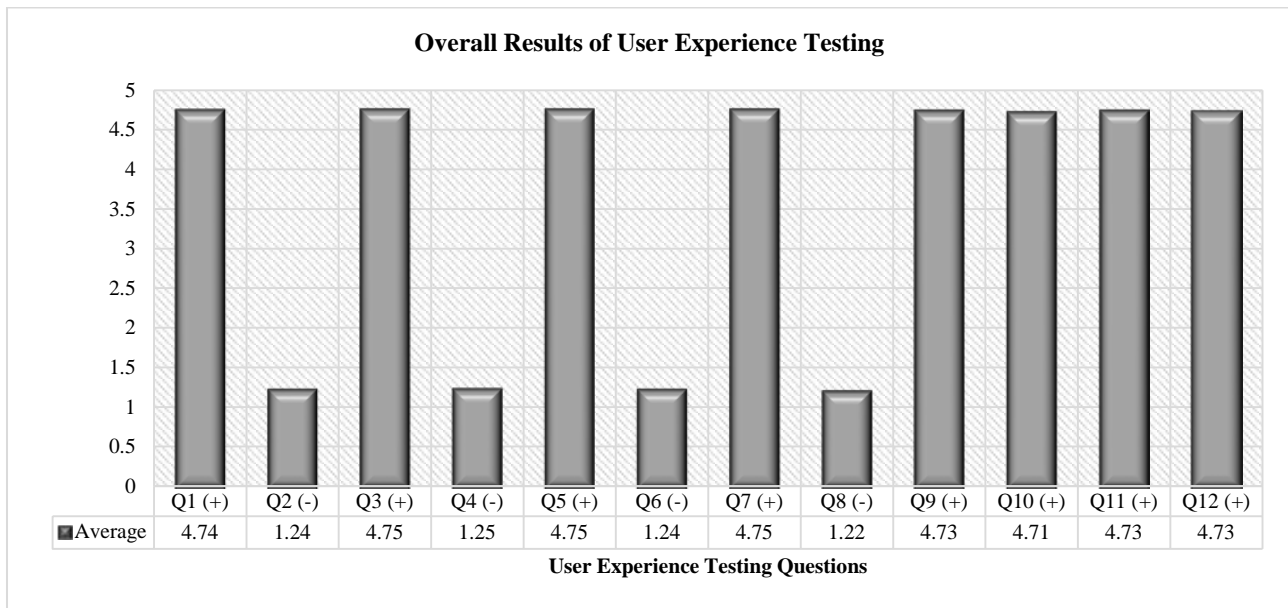


Fig. 5. Overall Results of user Experience Testing.

VII. CONCLUSION

Generally, the results from the testing indicate the success of the study. The findings from the positive-type-question managed to achieve the total average score of 4.74 over 5 against the widely accepted score-5 in Likert-scale, which

suggested that users have good experience of using the conversational agent. The negative-type-question on the other hand managed to achieve the total average score of 1.24 over 5. In contrast with positive-type-question, the lower figure in Likert-scale for negative-type-question indicated that users somehow disagree that the usage of the conversation agent

constituent to bad experience. Apart from this, the test results from User Experience Testing indicated that 84.3% of the respondents were satisfied with the conversational agent, whereas 86% of respondents would use this conversational agent again. Thus, the results of the User Experience Testing demonstrated that users show positive experience towards the usage of the conversational agent. The analysis from User Experience Testing stipulated that most respondents were pleased with the conversational agent and would use it again. The respondents claimed that the chatbot was useful and they were able to retrieve answer quickly and effectively via this chatbot.

ACKNOWLEDGMENT

Special thanks to the Natural Language Computing Group, Microsoft Research Asia, Tunku Abdul Rahman University College (TAR UC), Universiti Teknikal Malaysia Melaka (UTeM) and MyBrain15 scholarship.

REFERENCES

- [1] Duijst, D., 2020. What is the effect of personalization on the UX of Chatbots? [online] Available at: <https://uxdesign.cc/what-is-the-effect-of-personalization-on-the-ux-of-chatbots> 392bf34bba3b/www.researchgate.net/publication/318404775 [Accessed on 5 May 2020].
- [2] Finstad, K., 2010. The Usability Metric for User Experience, *Interacting with Computers*, 22, pp. 323-327.
- [3] Davis, F. D. (1985). A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Massachusetts Institute of Technology.
- [4] Quarteroni, S., and Manandhar, S., 2008. Designing and Interactive Open-Domain Question Answering System, *Natural Language Engineering*, 1 (1), pp. 1-23.
- [5] Orlando, T.M. and Sunindyo, W. D., 2017. Designing dashboard visualization for heterogeneous stakeholders (case study: ITB central library), 2017 International Conference on Data and Software Engineering (ICoDSE), Palembang, 1-2 November 2017, pp. 1-6.
- [6] Rodriguez Gil, J. García-Zubia, P. Orduña, A. Villar-Martinez and D. López-De-Ipiña, 2018. New Approach for Conversational Agent Definition by Non-Programmers: A Visual Domain-Specific Language," *IEEE Access*, vol. 7, pp. 5262-5276.
- [7] Chancusing, J.C., and Bayona-Ore, S., 2019. Information and Communication Technologies Acceptance Models in Universities. 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, pp. 62-66.
- [8] Asastani, H.L., Harisno, V. H. Kusumawardhana and H. L. H. S. Warnars, 2018. Factors Affecting the Usage of Mobile Commerce using Technology Acceptance Model (TAM) and Unified Theory of Acceptance and Use of Technology (UTAUT), 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), Jakarta, Indonesia, pp. 322-328.
- [9] Setianto, F. and Suharjo, 2018, Analysis the Acceptance of Use for Document Management System Using Technology Acceptance Model, 2018 Third International Conference on Informatics and Computing (ICIC), Palembang, Indonesia, pp. 1-5.
- [10] Harb, Y., and Alhayajneh, S., 2019. Intention to use BI tools: Integrating technology acceptance model (TAM) and personality trait model, 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 494-497.
- [11] Natalia, S. Bianca and I. A. Pradipta, 2019. Analysis User Acceptance of Wonderful Indonesia Application Using Technology Acceptance Model (case study: Indonesian Ministry of Tourism), 2019 International Conference on Information Management and Technology (ICIMTech), Jakarta/Bali, Indonesia, pp. 234-238.
- [12] Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D., and Cochran, J.J., 2018. *Statistics for Business and Economics*, 13th Ed., Boston, USA: Cengage Learning.
- [13] Elsholz, E., Chamberlain, J., and Kruschwitz, U., 2019. Exploring Language Style in Chatbots to Increase Perceived Product Value and User Engagement, *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR)*, Glasgow, UK, 10-14 Mar 2019, pp. 301-305.
- [14] Carisi, M., Albarelli, A., and Luccio, F.L., 2019. Design and Implementation of an Airport Chatbot, *EAI International Conference on Smart Objects and Technologies for Social Good (GoodTechs '19)*, 25-27 September 2019, Valencia, Spain.
- [15] Chen, G., 2018. Nurturing the Companion ChatBot. 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), 2-3 February 2018, New Orleans, LA, USA.
- [16] Chung, M., Ko, E., Joung, H., and Kim, S.J., 2018. Chatbot e-service and Customer Satisfaction Regarding Luxury Brands, *Journal of Business Research*.
- [17] Wei, C., Yu, Z., and Fong, S., 2018. How to Build a Chatbot Framework and its Capabilities, *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, Macau, China, February 2018.
- [18] Tegos, S., Demetriadis, S., Papadopoulos, P.M., and Weinberger, A., 2016. Conversational Agents for Academically Productive Talk: A Comparison of Directed and Undirected Agent Interventions. *International Journal of Computer-Supported Collaborative Learning*, 11, pp. 417-440.
- [19] Weiers, R.M., 2011. *Introductory Business Statistics*, 7th Ed., Boston, USA: Cengage Learning.

Survey of Tools and Techniques for Sentiment Analysis of Social Networking Data

Sangeeta Rani¹, Nasib Singh Gill², Preeti Gulia³

Department of Computer Science and Applications
Maharishi Dayanand University, Rohtak, India

Abstract—Social media has rapidly expanded over a period of time and generated a huge repository of content. Sentiment analysis of this data has a vast scope in decision support and attracted many researchers to explore various possibilities for technique enhancement and accuracy improvement. Twitter is one of the social media platforms that are widely explored in the area of sentiment analysis. This paper presents a systematic survey related to Social Networking Sites Sentiment Analysis and mainly focus on Twitter sentiment analysis. The paper explores and identifies the techniques and tools used in a well-structured approach to find out the research gaps and identify future scope in this area of research. The techniques evolved over time to improve the efficiency of classification. Total 55 research papers are included in this survey. The result reflects that Twitter is the most explored social networking site for opinion mining. Naive Bayes and SVM machine learning algorithms are implemented in maximum researches. As the latest advancements, Stack based ensemble, fuzzy based and neural network based classifiers are also implemented to enhance the efficiency of classification. WEKA, R Studio, Python are mostly used tools by research scholars for implementation. The overall evolution of the research goes through various changes in terms of technologies, tools, social media platforms and data corpus targeted.

Keywords—Social networking sites sentiment analysis; twitter sentiment analysis; opinion mining; ensemble classifier; stack based ensemble

I. INTRODUCTION

The spread of information on social networking media like Facebook, Twitter, Instagram, Reddit, News forum etc. is comparatively faster than traditional social media platforms. Social media have become a rich resource of information for companies and research scholars that can be analyzed to get valuable information by using NLP (Natural Language Processing) and artificial intelligence techniques. The huge repository of information provided on social media platform is unprocessed and raw in nature, and over the time technologies are evolved to process the data and extract valuable information from that. This information can be analyzed and helpful in decision support and effective policy making in different areas related to business, politics, entertainment, medical and social uplifting.

Sentiment analysis of social media posts deals with finding out the opinion, sentiment or feelings related to these posts.

That can be mentioned at different levels of sentiments and mostly categorized as positive and negative. Several sentiment analyses and classification techniques like dictionary based, machine learning, ensemble based, neural network based, fuzzy based and hybrid are evolved over the period of time starting from the research in the area. Also, the targeted data size is increased and new tools are evolved for easy and effective evaluation of sentiment. Various research scholars have been doing research for more than a decade and research has gone through multiple phases with enhancement of technology and efficiency of outcomes.

Here in the present survey we have gone through a systematic literature survey and studied 55 finally selected research papers related to the area from 2009 to 2021. These 55 papers are selected after keen observation and following the criteria of inclusion and exclusion. We focus on Twitter sentiment analysis and provide the existing techniques used and scope of enhancement. There is abundance of research literature present in the field; we aim to find the relevant literature with respect to novelty of research, their applications domain and effectiveness.

Section I of the present survey paper gives the introduction of sentiment analysis for social networking sites. The research strategy used in the survey is mentioned in Section II. Research questions on the basis of which the survey is designed are mentioned in Section III. Section IV gives the details of related literature included in the survey. Survey outcomes of all the 55 research papers included are mentioned in Section V. Overall survey is concluded in Section VI.

II. RESEARCH STRATEGY DESIGN

The survey related to ‘Social networking sites sentiment analysis’ was undertaken systematically by following the steps mentioned in Fig. 1. At the very first step, research questions are designed to give a proper direction to the survey. We continue by retrieving the related literature and then selecting the pertinent research papers from those that fulfill the requirement as per research questions. Finally, the findings and results as started by the author are analyzed and mentioned along with tools and technology used in the research.

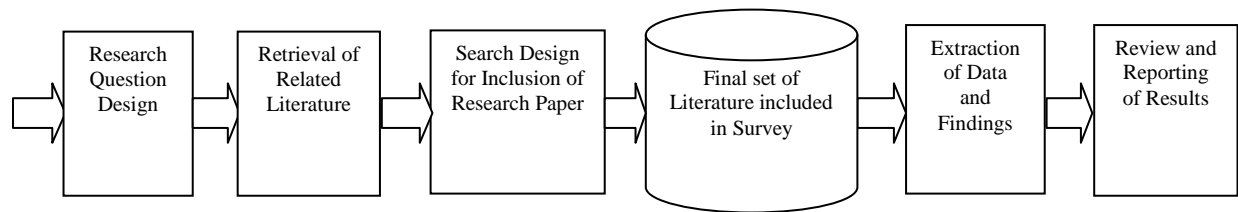


Fig. 1. Stepwise Strategy Design for Literature Review.

III. RESEARCH QUESTION

To design and conduct research, the following research questions (Q) are identified:

- Q1: Which technique is used in the research?
- Q2: Which framework, language or tools are used in research for implementation?
- Q3: Which datasets are targeted in the research for implementation of mentioned models?
- Q4: What are the main outcomes of the research literature?

IV. RELATED LITERATURE

Various researches performed in the area from 2009 to 2021 are mentioned below. Table I gives the summarization and key findings of the survey. The results, technologies, findings and tool used in the research are also elaborated in the section below:

Matthew et al. [1] in their research, implemented Bagging, Boosting and Random subspace method by using KNN, C4.5, SVM, MLP, RBF, LR as base classifiers. WEKA tool is used in the implementation of different classifiers. An enhanced performance is obtained with maximum accuracy of 90% over other approaches. Ensemble-based classifier performed better in all cases, particularly for noisy data, to enhance the overall accuracy of classification.

Kumar et al. [2] presented an article about the evolution of online social networks. The article investigates the dynamics of social cognitive theory and social networking. The research involved a photo sharing application (Flickr) and Yahoo 360 social network for analysis and implementation. Three different segments of networks are identified viz. singletons, isolated communities and giant component networks and detailed description of the evolution and structure of the three segments are researched on. The investigation of economical behavior of online social media networks is analyzed and user activity impact on incentives is examined.

A. Agarwal et al. [3] proposed twitter sentiment analysis by using POS specific polarity features and explored tree kernels to prevent the need for tedious feature engineering. 11,875 manually labeled tweets publically available from commercial resources were used in implementation. The conjunction of new feature with the previously proposed features and tree kernels outperform the base line classifiers.

Bae Y. et al. [4] performed twitter sentiment analysis of twitter post related to famous personalities and news channels viz. Donald J. Trump, Barack Obama, Bill Gates, Ashton Kutcher, Lady Gaga, Larry King, Oprah Winfrey,

Britney Spears, TechCrunch, Mashable, BBC Breaking News, CNN Breaking News, Dalai Lama.

Lima et al. [5] in their research implemented Naïve Bayes algorithm for tweets sentiment classification as positive and negative on real time tweets downloaded by using Twitter4J Library. Tweets are classified on the basis of emoticons, sentiment based words or a hybrid of both. The results show an enhanced accuracy in case of hybrid approach.

F. Neri et al. [6] implemented sentiment analysis on Facebook posts related to news post related to Rai1 and La7 news programs. Facebook posts are analyzed using 'iSyn Semantic Center'. Bayesian method and K-Means algorithm are used as supervised and unsupervised classification techniques. The research shows the importance of Facebook for online marketing.

H. Kang et al. [7] proposed an enhanced Naïve Bayes classification algorithm for sentiment classification of review documents of restaurants. The 70000 review documents are obtained from restaurant sites including star information. The proposed model shows the enhancement of accuracy and precision.

M. Ghiassi et al. [8] developed a new lexicon specifically for Twitter opinion mining using n-gram feature vector and supervised learning method. The 3440 tweets are manually collected and labeled on 'Justin Bieber' twitter account and the model proposed in research is tested using these tweets. The results show the improvement in accuracy for the proposed model over SVM with an accuracy of 95.1%.

Hassan et al. [9] implemented Bootstrap ensemble framework (BPEF). It works in two stages: expansion and contraction. In the expansion stage, large numbers of models are generated based on the dataset, features and classifier parameters. In contraction stage a subset of these models is selected by throwing redundant and less useful models. The experiment results show that BPEF gives a high value of recall as compared to other methods. SIMS module of BPEF extracted a model with higher performance.

E. Haddia et al. [10] show the role of preprocessing on sentiment analysis. Improvement is observed in the accuracies of TD-IDF matrix from 78.33 to 81.5, in Metric FF 76.33 to 83 and in FP matrix 82.33 to 83.

Patil et al. [11] in their research implemented SVM with and without feature extraction and show that SVM eliminated the need for feature selection due to the ability to generalize high dimension feature space.

Inoshika et al. [12] research on feature ranking and selection techniques for Twitter data opinion mining and

suggested to remove unrelated words from feature space to reduce dimensionality that further reduces the sparseness of the feature set. The research also proposed a new feature selection technique on the basis of information theory named as Ratio Method.

Bac Le et al. [13] proposed a model based on NB and SVM. Information Gain, Bigram, Object-oriented extraction methods are used for feature ranking and selection to select more appropriate features. As per reported results, the proposed model is highly efficient with high accuracy for predicting feelings.

B. S. Dattu [14] implemented twitter sentiment analysis by using SVM and Naive Bayes on real time tweets downloaded between the time periods 12 September 2010 to 24 January 2011 with the keywords 'NFL teams'. They pointed out in their research that SVM proved to be better than Naive Bayes algorithm for text classification and categorization. For unbalanced data, Naïve Bayes is more appropriate as there are fewer variations in results for unbalanced data.

O. Kolchyna et al. [15] implemented two techniques viz. Lexicon based and machine learning for sentiment classification of twitter messages. This research uses the sentiment score extracted from Lexicon classifier as an additional feature in the feature vector and the results shows the improvement in accuracy for imbalanced data set. The research show that incorporating sentiment lexicons with abbreviations, emoticons and social media slang enhances the efficiency of lexicon-based classifier. Feature generation and selection also play a vital role for the enhancement of classification accuracy. SemEval-2013 competition, task 2-B standard twitter data set is used and the outcome of the research shows that SVM and NB machine learning methods perform better. A combination of lexicon and machine learning method further enhance the accuracy by 7 percent.

Prusa et al. [16] worked on bagging and boosting-based ensemble classifiers. These two are the most widely used ensemble techniques in machine learning. In the research, both techniques are tested with the use of seven diverse base learners. All the ensemble classifiers build are compared with all the seven base learners to observe the performance enhancement. Total 21 learning algorithms are trained and finally tested on two different datasets, one large-sized automatically class labeled lesser quality dataset and other small-sized manually class labeled superior quality dataset. The research proved to be better and ensemble classifier enhanced the accuracy, regardless of the quality of the data set used.

K. L. Devi et al. [17] compared ensemble classifiers viz. Boosting and Bagging with the machine learning classifiers like NB, SVM and maximum entropy classifier. Feature selection is performed by using MI and Chi-square methods and is proved to be better than previously used methods. SemEval 2013, Task 9 data sets are used for implementation.

Y. Wan et al. [18] in their research, implemented majority voting-based ensemble classification model on various classification techniques including SVM, Random Forest, Naive Bayes, Bayesian Network and C4.5 Decision Tree by

using 10 fold cross validation on a data set having 12864 tweets related to airline service Twitter dataset.

Prusa J. et al. [19] researched on 10 different feature selection methods and four classifiers viz. 5-NN, C4.5, LR, MLP. All classifiers by using all feature ranking and selection techniques are implemented on 10 different sized feature subsets up to maximum 200 features. The results of research show that filter-based feature selection, Chi-Squared (CS) improved classification performance for small-size feature sets. After comparing all the combinations of classifiers and feature rankers, it was observed that LR performed best with 150 features selected by KS ranker. The models performed better with larger number of features and the best models have features 75 or more. Only the feature rankers MI, ROC, PRC, CS and KS shows enhanced efficiency as compared to no feature selection.

R. Mansour et al. [20] in their research used multiple sets of features for sentiment classification by using an ensemble classifier. The classification complexity comes out linear with the increase in feature set size. The ensemble is implemented on two features sets; one optimal set with 20000 features and other NRC data set with 4 million features. The feature set with selected 20000 features has shown relative 9.9% and 11.9% performance gain over 4 million feature set.

O. Abdelwahab et al. [21] in their research demonstrated the effect of training set size on accuracy of SVM and NB classifiers. The Python NLTK library is used for implementation of classifiers. The results show that there is a little increase in accuracy if training data increases from 20 to 90 to percent. So a moderate size data can be trained to get acceptable results.

S. Akter et al. [22] predicted the sentiment for the Facebook posts using a lexicon-based sentiment analysis technique. The data set used in the implementation is FOODBANK the Facebook group in Bangladesh. In the research a console is developed using C# and Graph API is used to collect data.

M. Bouazizi et al. [23] proposed a new model for detecting the sarcasm using sentiment analysis of twitter data as micro blogging social networking sites are very useful in detecting sarcastic statements. A pattern-based sarcasm detection approach is used for twitter. A feature set with four relevant features for identifying different kinds of sarcasm is used and tweets are classified as non-sarcastic and sarcastic two classes. The model achieved an accuracy of 83.1% with 91.1% precision. SVM classifier is used and WEKA tool is used in implementation.

Grandin and Adan et al. [24] proposed a model Piegas for the sentiment of Portugal tweets. The Naïve Bayes classifier is implemented by using JavaScript and Ruby on Rails are used for the development of the system. The main requirement of the model is to develop a system with good usability and high precision.

Nádia et al. [25] proposed a new semi-supervised approach to solve the problem of cost of getting supervised data for machine learning. Unsupervised information retrieved from the similarity matrix created from unlabeled data is used with

various classifiers in place of classified data. Similarity matrix can be used as a powerful knowledge extraction tool to get information from non-labeled data. The results of the proposed framework show the improved accuracy for Twitter sentiment classification by using unlabeled data.

A. Tripathy [26] implemented four machine learning classification algorithms SVM, NB, Maximum Entropy (ME) and Stochastic Gradient Descent (SGD) on IMDb data set for sentiment classification. These classification models are implemented on unigram, bigram and n-gram features and it is observed that if the value of n is increased in n-gram after 2 than accuracy is decreased rather than increasing. For unigram and bigram accuracy is good but for trigram, four-gram, five-gram accuracy is decreased. Also, the use of count vectorizer technique and TF-IDF for converting the text into a matrix of weights, enhance the accuracy of classification.

A. Krouska et al. [27] in their research show the effect of preprocessing on the classification accuracy. The research also shows a major enhancement in result when IG is used for attribute selection. The research uses Unigram, Bigram and 1-3 gram feature vector. Preprocessing and feature selection enhance the accuracy. Unigram and 1-3 gram performed best among all.

K. Ali et al. [28] proposed SAaaS (Sentiment Analysis as a Service) framework to abstract sentiments of various social media information services. Public health surveillance related to social media based is done by using spatial attributes of social media users to find the location of disease outbreak. A new quality model is introduced to remove noise from the social media content. The real-world datasets from Twitter, Instagram, Reddit, news forum are used in the research. Sentistrength and Alchemy API tools are used in research. The Sentistrength tool is used for the analysis of short and informal text while Alchemy API for long and formal text.

A. U. Hassan et al. [29] In this paper, presented the method to detect the depression level of a person by fetching emotions from the social media text by using NLP and machine learning techniques on social media Twitter dataset and 20 newsgroups.

M. R. Huq et al. [30] used two techniques for sentiment analysis: First technique is a sentiment classification algorithm (SCA) based on KNN and the second is based on SVM. The research shows the comparative analysis of both the methods on the basis of recall, precision, accuracy, F-Score, TPR and FPR for 1000 tweets.

M. Ahmad [31] implemented Support Vector Machine (SVM) on two twitter pre-classified data sets for textual polarity detection. Recall, Precision and F-Measure are used for comparative analysis. The result shows that performance of SVM depends on the dataset itself. So it can be an area of research that what kind of classification algorithm is good for which kind of data set and what is the reason for that.

J. Brandon et al. [32] implemented twitter sentiment analysis to find the opinion of people for candidates in 2016 US Presidential elections. Lexicon based classifier and NB Machine learning classifiers are used on two data sets. One data set is a manually labeled Twitter data set and the other is

an automatically labeled data set based on Hashtags and topic. A high correlation of 94 percent was found with polling data by using a moving average smoothing technique.

R. Wijayanti et al. [33] proposed an ensemble classifier based on a voting-based technique and used SVM, NB, LR (Logistic Regression) and Decision Tree classification algorithms for the implementation of proposed ensemble classifier. They used various feature representation techniques such as TF-IDF, sentiment lexicon score and term presence in their research. Ensemble classification results are proved to be better than individual machine learning classifiers, but ensemble accuracy highly depends on the selection of single classifiers used for creating the ensemble classifier.

Z. Jianqiang et al. [34] monitored the effect of six preprocessing techniques by using four classification algorithms (NB, SVM, RF, LR) and two feature selection methods. The result shows that accuracy is improved after using preprocessing techniques on the dataset. But removal of URL's, numbers and stop words hardly affects the accuracy, so they can be removed. Random deletion of words reduces the accuracy as the deleted word might be important in sentiment detection. NB and RF are more sensitive to the use of different pre-processing techniques.

Rahman et al.[35] analyzed the reliable decision making for a friend request to be accepted in Online Social Networks. Here, a quantitative study for analyzing the friend request has been carried out and the information regarding the social media websites were explained and information misuse of the other users and friends due to being deficient in trustworthy Friend Request Acceptance. In the research, a method is proposed for reliable friend request acceptance in Online Social Networks by finding out more details of the person who has sent the friend request.

Jianqiang et al. [36] proposed a method for opinion mining using deep convolution neural networks. Unsupervised learning is used for obtaining word embeddings by using a large set of Twitter data. The n-grams features combined with the word embeddings and polarity score extracted from sentiment lexicon are used for Twitter sentiment analysis. Sentiment classification labels were predicted after training the feature set with deep convolution network. GloVe-DCNN on the STSTd dataset performed best with accuracy 87.62%.

K. Tago et al. [37] performed an analysis based on Twitter data using user relationships and analysis of emotional behaviors. Here, two dictionaries of emotional words are analyzed using the machine learning classifiers and keyword matching is used for calculating emotion scores. Moreover, with different settings, three experiments were designed and these are the user's average emotion scores that were calculated. Using all the emotional tweets, the average of emotion score is calculated after user of few emotional tweets was excluded. Brunner-Munzel test was used to evaluate emotional behaviors to user relationships. As per results, positive users participate more than negative users in building a relationship in some particular conditions.

In Ikoro, Victoria, et al. [38], sentiment analysis of UK energy consumers is done by using messages posted on

Twitter. Big Six and three new entrant energy providers companies are compared on the basis of tweet sentiments. Two sentiment lexicons are used to maximize accuracy. As per results, consumers are more positive towards new companies and the use of multiple lexicons helps to improve the accuracy of sentiment analysis.

C. Troussas et al. [39] implemented four different ensemble techniques: bagging, Boosting, Voting and Stack based ensemble on three different data sets. Stack-based ensemble model is implemented by using NB, SVM, KNN and C4.5 as base classifiers and LR as a Meta classifier. The result shows that stack-based model surpasses the efficiency of other classifiers. Three datasets are used viz. OMD, HCR, STS-Gold. Stack based ensemble classifier performed best with an accuracy of 89.02% on STS-Gold dataset.

M. M. Fouad et al. [40] focused their research on the efficient classification of twitter data by combining NLP and data mining techniques. They implemented majority voting-based ensemble classification technique by using SVM, LR, NB classifiers and Information Gain (IG) feature selection technique. IG technique enhances the efficiency of classifier by selecting more appropriate features. The ensemble classifier also improves the accuracy, but if any one of the participant algorithms in ensemble does not suit the data set, then accuracy is decremented. Feature subset further enhanced with the use of emoticons does not enhance the efficiency of classification.

Y. Emre Isik et al. [41] also used stack-based ensemble classification techniques for sentiment classification of text. The ensemble is performed at two levels, one at feature selection level and other at classifier level; as a less accurate feature selection can lead to poor classification, so two techniques are used at feature selection level to reduce error and for enhanced feature selection. Two classification methods are used as an ensemble to enhance classification accuracy. The technique shows good results as compared to other machine learning classifiers.

F. T. Giuntini et al. [42] in their research perform the sentiment analysis on Facebook post. The aim of the research is to find the relevancy of emoticons used in the Facebook posts whether the emoticons match with the actual sentiment present in the post at six basic emotion levels. The paper proposed a 'Expectation Maximization algorithm' that finds correlation between the emoticons used in the tweet and the emotion class of the post. As per research, the use of emoticons as attribute enhances the result of classification of Twitter posts.

S. E. Saad et al. [43] used ordinal regression for twitter data sentiment analysis for twitter dataset provided by the NLTK. The algorithms used for opinion mining in the proposed model framework are SoftMax (Multinomial logistic regression), SVR (Support Vector Regression), DTs (Decision Trees), and RF (Random Forest). As per research, proposed framework can detect ordinal regression and decision tree proved to be the best from the above mentioned algorithms.

S. Vashishtha at al. [44] performed opinion mining of tweets by using three different lexicons and nine public twitter

data sets, for classifying tweets at two and three levels of sentiments. The research proposed a fusion of multiple lexicons with fuzzy classification approach for enhanced classification. The nine datasets used in implementation are viz. The dataset used are Sanders Twitter Dataset, Nuclear Twitter Dataset, Apple Twitter Dataset, (STS-Test), Sentiment140, SemEval 2017, SemEval 2015 and Data used by Gilbert & Hutto, 2014, SemEval 2016.

K. Elshakankery et al. [45] proposed a new hybrid approach named HILATSA by combining lexicon and machine learning approach. The proposed approach performed with an accuracy of 73.67 % for three class classification and 83.73 % for two class classification problem. Six different data sets used in the implementation are ASTD, Mini Arabic Tweets Sentiment Dataset, ArSAS, Arabic Gold Standard Twitter Data set, Syrian Tweets Corpus and Twitter dataset for Arabic Sentiment Analysis.

Martin-Domingo et al. [46] used machine learning classification for airport service quality analysis on London Heathrow airport's Twitter account dataset by using machine learning sentiment analysis technique. They used Theysay and Twinword tools for implementation. Theysay performed better than Twinword with 78.7 percent accuracy as compared to 69.6% of Theysay. The purpose of research is to generate a list of service attributes that reflect the ASQ and results reflect that additional attribute does not reflect more accurate ASQ prediction.

M. Naz at al. [47] in their research implemented an ensemble classification model by using two classifiers K-Nearest Neighbor and Naïve Bayes. They used two feature selection techniques: Forest Optimization algorithm (FOA) and minimum redundancy and maximum relevance (mRMR). FOA is used for feature selection and mRMR for the removal of irrelevant features. As per the results of the research, ensemble classifier combined with feature selection technique performed comparatively better than the individual machine learning algorithms. Results are further improved by using an ensemble of KNN, NB and SVM. It is also evident from the research that the hybrid of FOAKNN and FOA-NB has outperformed single KNN and NB classifiers. Accuracy is increased when FOA and nRMR feature selection techniques are applied. The Blitzer's dataset, retrieved from the UCI repository related to the reviews of electronic products, is used for the implementation of various classifiers in the research.

J. J. Bird et al. [48] proposed multiclass sentiment classification for five different levels of sentiments in a range from 1-5 representing negative to positive score. Various single classifiers viz. OneR, MLP, NB, NBM, RT, J48, SMO SVM and ensemble classifiers named RF, AdaBoost (RT), Vote (RT, NBM, MLP), Vote (RF, MLP, NBM), AdaBoost (RF) are implemented. The research shows that the majority voting-based ensemble of NBM, RF, MLP performed best with 91.02 % accuracy. In individual classifiers, RT (Random Tree) performed best with accuracy of 78.6%. All ensemble methods outperformed single classifier.

M. Khader et al. [49] in their research show the effect of preprocessing techniques such as using tokenization, PoS tagging, removing stop words, URL, other users' mention,

numbers and hashtags and lemmatization on Naïve Bayes machine learning classifier. Mapreduce of Hadoop is used for the implementation on Stanford twitter Sentiment data set. The proposed technique reflects an increase of 5% accuracy yielding to 73% for NB classifier.

R. Ahujaa et al. [50] implemented six classifiers viz. Decision Tree, SVM, KNN RF, LR TF-IDF, NB by using two feature selection techniques N-gram and TF-IDF on ‘SS-Tweets’ data set. The results show that TF-IDF feature selection show 3-4 % increase in performance as compared to N-gram feature.

M. bibi et al. [51] in their research proposed a new feature selection technique CAARIA “class association and attribute relevancy based imputation algorithm” that is proved to be better than IG and PC with an AUC (F-measure) value of 0.79. The research is performed on three twitter data sets HCR, SS-Tweet and FleTweetsPak on two machine learning classifiers SVM and NB by using WEKA tool. The newly proposed technique reduces feature dimension space by selecting tweets that have same class and carry useful information.

M. Bibi et al. [52] used hierarchical based clustering techniques named SL (single linkage), AL (average linkage) and CL (complete linkage) for the sentiment mining of twitter data. A combined framework architecture is built by using these three clustering techniques to select the best possible cluster with the help of using majority voting. The hierarchical clustering techniques proposed in the research are compared with k-means, SVM and NB classifiers. The outcome of research indicates that majority voting-based cooperative clustering is better in terms of quality of clusters but poor in term of time efficiency.

Z. Kermani et al. [53] used IDF, Term Frequency, sentiment scoring using lexicon dictionary SentiWordNet, semantic similarity for representing each feature weight of tweet in the feature vector. The percentage of contribution in the weight by each method is optimized and solved by genetic

algorithm. The weight of feature obtained from all the four techniques are merged by using Einstein sum. SVM and multinomial NB classification methods are used on this weighted enhanced feature vector to classify tweets. Four Twitter data sets used for the implementation are Stanford testing dataset, Strict Obama McCain Debate dataset, STS-Gold and Obama-McCain Debate dataset.

Esraa A. Afify et al. [54] in their research aim to classify the Facebook account as fake or genuine, on the basis of content generated and finding the correlation between user generated content. Credibility of an account is decided at two levels. First binary classification is applied to classify account as fake or genuine. After that, credibility score of the genuine class is calculated by using Analytical Hierarchical Process. On the basis of that score account credibility is decided. The research used machine learning and deep learning techniques for the identification of Facebook profile credibility by using Scikit-learn and Keras with TensorFlow. Scikit-learn is used for the implementation of machine learning techniques and Keras with TensorFlow for implementing deep learning.

George S.R. et al. [55] proposed a framework for opinion prediction for a product or brand name in Facebook during social distancing by using machine learning algorithms and netnography. The study actually proposes a conceptual framework and suggested various tools used by different researchers for opinion mining of Facebook viz. netnography, Google analytics, tweetstats, brandwatch, Facebook insights, sematrix’s lexalytics, Google alerts and people browser.

Most of the work aim to find out the sentiment related to service or product by using dictionary-based, Machine learning based or hybrid classifiers. The overall purpose is to enhance accuracy and efficiency of classification models. Different researches are performed on different data corpus. Different tools are used by researchers to observe the variations in the outcome. Table I summarizes the tool, technology, data corpus used and final outcome of all the papers included in the survey.

TABLE I. SUMMARY OF REVIEWED LITERATURE

Author (Year)	Technique/ Approach	Tool Used	Data Corpus / Context	Results / Outcomes
M Whitehead et al. [1] (2009)	Ensemble classifiers: Bagging and AdaBoost-r (KNN, C4.5,SVM, MLP, RBF, LR)	WEKA	Twitter	Ensemble classifier performed better, for noisy data also.
Kumar et al. [2] (2010)	Evolution and analysis of Social Networks in three different size networks by Graph Theory.	Representation and analysis of social network as timegraph.	Yahoo!s 360 social network and Flickr photo sharing application.	Mention the evolution of social network components of different sizes and shows "star" as most prevailing structure of social network.
A Agarwal et al. [3] (2011)	Tree kernels with POS specific polarity features	Python	11,875 manually labeled tweets	Maximum tweets are classified in neutral class.
Bae Y. et al. [4] (2011)	Dictionary based classification using LIWC2007	Twitter API, LIWC2007	Twitter	Sentiment analysis of twitter data related to famous personalities as negative or positive.
ACES Lima et al. [5] (2012)	Naïve Bayes	Twitter4J Library , JAVA	Twitter Dataset: Real Time Tweets.	Sentiment classification on the basis of emoticons, sentiment based words or hybrid. Hybrid approach enhances accuracy.
F Neri et al. [6] (2012)	Bayesian method and K-Means algorithm	iSyn Semantic Center	Facebook about newscasts (La7 and Rai1 news programs)	Show importance of Facebook for online marketing.

H Kang et al. [7] (2012)	Naïve Bayes, SVM	WEKA	70000 review documents from restaurant sites	Proposed enhanced NB algorithm. Enhancement in precision and accuracy in proposed algorithm.
M Ghiassi et al. [8] (2013)	Supervised learning technique using n-gram statistical analysis	SVM and DAN2	Real time collected for keyword 'Justin Bieber'	Proposed model performed better than SVM with an accuracy of 95.1%.
Hassan et al. [9] (2013)	Bootstrap ensemble framework (BPEF).	FRN, Viralheat, Popular Tool, Light side, Lymbix, Sentistrength, Sentiment 140	Twitter dataset: Telco, Tech, Pharma	SIMS module of BPEF was able to extract models with higher efficiency.
E Haddia et al. [10] (2013)	Preprocessing Methods. Classifier SVM . Feature wt. using TD-IDF, FF, FP.	'e1071' in R Tool	movie reviews, DAT-1400, Dat-2000	Improvement observed in accuracies of TD-IDF matrix from 78.33 to 81.5, in Metric FF 76.33 to 83 and in FP matrix 82.33 to 83.
G Patil et. al [11] (2014)	ANN, SVM And TF-IDF	----	Twitter	SVM Performed better than ANN. SVM use few irrelevant feature and High dim. Feature space.
Inoshika et al. [12] (2014)	Machine Learning: SVM, Decision Tree, RF, J48, CART. Feature Selection: feed forward and feedback selection, IG , Chi square	----	Twitter Data set: Data by Ada Derana, News First, Ceylon Today, Lanka Breaking News, ITN.	Proposed a new information theory based feature ranker named Ratio Method that result in improved efficiency.
Bac Le et al. [13] (2015)	Machine Learning :SVM and NB IG, Bigram, Object-oriented extraction for feature selection	AlchemyAPI, Scikit-Learn tool	Twitter : AlchemyAPI, Zemanta, OpenCalais	Accuracy enhanced.
BS Dattu et al. [14] (2015)	SVM , NB	TwitterSentiment and SentiStrength	Real Time Twitter Data related to 'NFL teams'	SVM performed better than NB and NB is insensitive to unbalanced data.
O Kolchyna et al. [15], (2015)	Lexicon-based and machine learning (SVM, NB)	WEKA	SemEval-2013 competition, task 2-B	Combining lexicon with SVM and NB enhance accuracy.
J Prusa et al. [16] (2015)	Ensemble Technique : Bagging and Boosting using 7 base learners (5NN, C4.5D, C4.5N, MLP, LR, SVM, RBF).	WEKA	Sentiment140 (S_3000, S_359)	Ensemble Classifier enhances accuracy. Enhancement in result is less uniform for smaller size data set.
K.L Devi et al. [17] (2015)	Ensemble classifiers: Bagging and Boosting (SVM, NB, ME) Feature Selection using MI and Chi-square	----	Twitter dataset: SemEval 2013, Task 9	Ensemble based learners produce improve accuracy than base learners.
Y Wan et al. [18] (2015)	Majority voting ensemble(NB, SVM, Bayesian Network, RF and C4.5) and IG Feature selection	WEKA	12864 tweets related to airline service Twitter dataset.	Ensemble outperformed with an accuracy of 91.7%. RF learner was best in machine learners with accuracy 90.8%.
J Prusa et al. [19] (2015)	Four classifiers(5-NN, C4.5, LR, MLP) 10 feature selection methods: CS, GI, KS, MI, PR, PRC, ROC, S2N, SAM, WRS.	WEKA	Twitter Dataset: Sentiment140	Best feature selection methods - CS and MI. Best feature set size - 100 to 200. CS, MI, PRC, KS and ROC resulted in performance enhancement in comparison to using no feature selection
R Mansour et al. [20] (2015)	Ensemble Classification. Feature selection. NCR Feature set, LLR Feature selection.	Natural Language Toolkit (NLTK)	Twitter Dataset: SemEval and CrowdScale	Performance gain of 9.9% on CrowdScale and 11.9% on SemEval.
O Abdelwahab et al. [21] (2015)	SVM and NB	Python NLTK library	SEMEVAL 2014	SVM is more Robust than NB but NB is faster comparatively. Increase in training set size after 20 % have very little effect on accuracy enhancement.
S Akter et al. [22] (2016)	Lexicon Based Approach / Dictionary Based Approach.	Graph API, C#	FOODBANK: Facebook group in Bangladesh	Identified sentiment behind a status post of Facebook by using lexicon based approach.
M Bouazizi et al. [23] (2016)	SVM	WEKA and OpenNLP, Gate Twitter partof-speech tagger,	Twitter	Sarcasm Detection: Accuracy of 83.1% with to 91.1% precision.
P Grandin et al. [24] (2016)	NB	JavaScript and Ruby on Rails	Twitter	Sentiment classification of Portugal tweets, Good Accuracy with improved classification time.
Nádia et al. [25] (2016)	Hybrid (Classification and Clustering Ensembles) : Semi-supervised approach (Labeled and Unlabeled data), SVM, Lexicon Based	----	SMS2013, Twitter2014, Twitter Sarcasm 2014, LiveJournal, Twitter2013.	Proposed framework shows the improved accuracy (80%).

A Tripathy et al. [26] (2016)	SVM, NB, Maximum Entropy and Stochastic Gradient Descent. Features: unigram, bigram, Trigram and n-gram features.	----	Twitter IMDb data set	Up to bigram feature set accuracy enhances. After that for tri, four or five gram accuracy of decreases. A combination of Bi gram, Tri gram with POS also worked well. Unigram + Bigram (ME) - 88.42 Uni + Bi + Trigram (ME) - 83.36 Uni + Bigram (SVM) - 88.884 Uni+Bi +Trigram (SVM) - 88.944 Bigram (SGD) – 95
A Krouska et al. [27] (2016)	Preprocessing Methods. IG Feature Selection. Uni, Bi and 1-3 gram features. Machine Learners – NB, KNN, SVM, C4.5.	Snowball stemmer library, Rainbow list	Twitter OMD HCR STS-GOLD	Feature extraction improve accuracy. unigram and 1-to-3-grams perform better.
K Ali et al. [28] (2017)	SAaaS Model using spatio-temporal properties and Lexicon classifier	Sentistrength and Alchemy API . .Net framework by using ASP.Net/C#	Real world datasets from twitter, Instagram, Reddit, news forum	Identified sentiment and location of disease outbreak.
AU Hassan et al. [29](2017)	Machine Learning classifiers(SVM, NB, Maximum Entropy) Binary and Multiclass classification, Voting	----	twitter dataset and 20newsgroups	Detect the depression level of users. SVM performed better than NB. SVM - 91%, NB – 83%, ME – 80%.
MR Huq et al. [30] (2017)	k-nearest neighbor, SVM	Java language	Twitter	KNN with normalization and keyword base(5 features) performed good.
M Ahmad et al. [31] (2017)	SVM	WEKA	Twitter Dataset related to self-driving cars and apple products	Performance of SVM highly depends on Dataset.
J Brandon et al. [32] (2017)	Lexicon-based and machine learning	National Language Toolkit (NLTK)	Twitter	High correlation of 94 percent was found with polling data.
R Wijayanti et al. [33] (2017)	Voting based ensemble technique (NB, SVM, decision tree, LR) and feature selection using TF-IDF, sentiment lexicon score and term presence	----	Twitter: Indonesian Twitter messages, Indonesian online marketplaces	Ensemble classifier enhance efficiency (accuracy - 91.59% and F1-score - 91.59%) but highly depend on base classifiers.
Z Jianqiang et al. [34] (2017)	Six Preprocessing Methods. Machine Learning Classifiers – NB, LR, SVM, RF.	GridSearch to find optimal parameters. scikit-learn for classifier.	Twitter Dataset: STS-Gold, SE-Twitter, SemEval2014, SS-Twitter, STS-test.	Removal of URL's, stop words and numbers hardly affects the accuracy. Random deletion of word reduces the accuracy. NB and RF are more sensitive to use of different pre processing techniques.
Rahman et al. [35] (2018)	Tree based algorithm, Depth search based algorithms.	T-Test	Facebook Accounts of UMP students	Detect reliable friend request.
Jianqiang et al. [36] (2018)	Deep convolution neural networks. BoW-SVM and BoW-LR	----	Twitter Data Sets: STSTd, SemEval2014 Task9, STSGd, SED, SSTd	GloVe-DCNN on the STSTd dataset performed best with accuracy 87.62%.
K Tago et al. [37] (2018)	Machine learning	T Test, Brunner–Munzel test	Twitter: Five thousand twitter accounts by using Twitter API.	As per results positive user participate more than negative users in building relationship in some particular conditions.
V Ikoru, et al. [38] (2018)	Lexicon Based	----	Twitter : Tweets of UK energy consumers	Consumers are more positive towards new companies and Multiple lexicon help to improve accuracy of sentiment analysis.
C Troussas et al. [39] (2018)	Voting base ensemble, Bagging and Boosting, Stack base ensemble (NB, SVM, KNN and C4.5) LR as Meta classifier	WEKA	Twitter OMD, HCR, STS-Gold	Stack base ensemble performed best with accuracy 89.02% in case of STS-Gold dataset.
MM Fouad et al. [40] (2018)	Majority voting based ensemble technique using SVM, LR, NB classifiers. IG feature selection	Java (Stanford Core NLP library) for feature extraction RapidMiner	Stanford-1K, Stanford-3K, Sanders, HCR	Ensemble classifier enhances efficiency but highly depend on base classifiers. IG selection boosted accuracy.

Y Emre Isik et al. [41] (2018)	Stack base ensemble classification, Ensemble at feature selection level.	----	Twitter	Enhanced accuracy than machine learners.
FT Giuntini et al. [42] (2019)	Expectation Maximization algorithm	Lime Survey tool,	Facebook	Relevancy of emoticons used in the Facebook posts
SE Saad et al. [43] (2019)	SoftMax (Multinomial logistic regression), SVR(Support Vector Regression), DTs(Decision Trees), and RF (Random Forest)	Python software	Twitter dataset provided by the NLTK	Decision tree proved to be the best.
S Vashishtha at al. [44] (2019)	Unsupervised Fuzzy Classification, SVM. Lexicon: SentiWordNet, AFINN, AFINN	Python	Twitter Dataset : Sanders Twitter Dataset, Nuclear Twitter Dataset, Apple Twitter Dataset, (STS-Test), Sentiment140, SemEval 2017, SemEval 2015 and Data used by Gilbert & Hutto, 2014, SemEval 2016.	The method based on Fuzzy Rule performed better than SVM. VADER and AFINN lexicon outperformed as compared to SentiWordNet.
K Elshakankery et al. [45] (2019)	HILATSA as Combination of lexicon based approach and machine learning classifiers (SVM, L2 Logistic Regression, RNN)	JAVA Libraries- LIBSVM, DL4J and LIBLINEAR	ASTD, Mini Arabic Sentiment Tweets Dataset, ArSAS, Arabic Gold Standard Twitter Data set, Syrian Tweets Corpus and Twitter dataset	The proposed approach with accuracy: 73.67% - 3-class classification 83.73% - 2-class classification
Martin-Domingo et al. [46] (2019)	Machine learning	Twitter Archive, Theysay and Twinword	Twitter : London Heathrow airport's Twitter account dataset	Theysay performed better than Twinword with 78.7 percent
M Naz at al. [47] et al. (2019)	Ensemble Technique using K-Nearest neighbor NB and SVM. FOA and mRMR feature selection algorithm.	MATLAB	Twitter : Blitzer's dataset, retrieved from UCI repository	Ensemble of KNN, NB, SVM with feature selection enhances accuracy.
JJ Bird et al. [48] (2019)	OneR, MLP, NB, NBM, RT, J48, SMO SVM, Ensemble classifiers: RF, Vote (RF, MLP, NBM), AdaBoost (RT), AdaBoost (RF), Vote (RT, NBM, MLP).	JAVA	Twitter : London based restaurant tweets from TripAdvisor	Vote (RF, NBM, MLP) performed best with 91.02 % accuracy.
M Khader et al. [49] (2019)	Machine Learning Classifier: Naive Bayes PoS Tagging, lemmatization, weighting terms.	Apache OpenNLP of MapReduce (Hadoop).	Stanford Twitter Sentiment data set	5% increase in accuracy for Stanford Sentiment data set.
R Ahujaa et al. [50] (2019)	Machine Learning Classifier: Decision Tree, SVM, KNN RF, LR TF-IDF, NB. Feature selection: N-gram TF-IDF	----	Twitter Dataset: SS-Tweets	TF-IDF performance is 3-4 % higher than N-gram feature.
M bibi et al. [51] (2019)	Machine Learning: SVM, NB Feature Selection : IG and PC	WEKA	Twitter Dataset: HCR, SS-Tweet, FleTweetsPak	CAARIA: Proposed feature selection technique. CAARIA proved to be better than IG and PC with AUC (F-measure) value 0.79.
M Bibi et al. [52] (2020)	Clustering and classification technique using majority voting, k-means, SVM and NB classifiers	WEKA	Twitter Data Sets : HCR - Health Care Reform, SS-Tweet - Sentiment Strength Twitter Dataset STS-Test - Stanford Twitter Sentiment Test Set	Majority voting based cooperative clustering is better in terms of quality
Z Kermani et al. [53] (2020)	Machine Learning-based approach and genetic algorithm for calculating feature weight.	NLTK, Scikit packages of Python software	Twitter Data Sets : Stanford test data corpus, STS-Gold, Strict Obama McCain Debate, Obama McCain Debate Datasets.	Efficiency of proposed method TSA is improved but time complexity is poor, specially for big twitter data sets.
EA Afify et al. [54] (2020)	Machine Learning, Deep Learning.	Machine Learning-Scikit-learn, Deep Learning-Keras with TensorFlow	Facebook	Evaluate Facebook profile credibility. Identify fake and genuine users.
SR George et al. [55] (2021)	Machine Learning Classifier.	Netnography	Facebook	Proposed conceptual framework for Facebook opinion mining

V. SURVEY OUTCOMES

In the present survey paper we have conducted a systematic survey of literature related to social networking site's sentiment analysis or opinion mining. From 2009 to 2021 several researches have been conducted and technology evolved from simple dictionary-based sentiment prediction to ensemble, fuzzy, deep learning and neural based sentiment analysis. Detailed outcomes of researches are mentioned in Table I with tools and technologies used. Various advancements occur in the area of attribute selection, preprocessing and classifiers used. Data becomes big and technology changed as per data need. From the detailed survey of included literature, it has been observed that Naïve Bayes and SVM are the most explored machine learning classifiers. Lots of work has been conducted in the area of ensemble classification technique. The most of the researchers are attracted by Twitter opinion mining and Facebook is the second most explored social media platform. Sentiment 140 is quite a frequently used data corpus. WEKA, RStudio, Python and NLTK are used in several research implementations. Facebook and Twitter are relatively less unstructured and sentiment analysis does not include image, audio or video. As media content can be in any one of these forms also, so sentiment extraction from these resources can be quite interesting and important but challenging too. So lots of work has been done on text sentiment analysis and most of them target to improve efficiency of classification.

VI. CONCLUSION

The manuscript presents a survey conducted on 55 different research papers related to social networking site's sentiment analysis. The survey reflected the evolution and enhancement of tools and technologies from 2009 to 2021 for sentiment analysis. Twitter is the maximum explored social networking site in the area of sentiment analysis. WEKA, RStudio and NLTK are most popular tools used by researchers. The area of text sentiment classification has been widely explored with the use of advanced classification techniques, big data technology, better simulation tools and most of them target to improve efficiency of sentiment classification. A new scope can be sentiment analysis from images, audio and video content as this area is comparatively untouched and huge repository of audio and video content is available on social media.

REFERENCES

- [1] M. Whitehead and L. Yaeger, "Sentiment Mining Using Ensemble Classification Models", Springer "Innovations and Advances in Computer Sciences and Engineering", Dec 2009, pp. 509–514.
- [2] R. Kumar, J. Novak, A. Tomkins, "Structure and evolution of online social networks", in 'Link mining: models, algorithms, and applications', Springer, pp. 337– 357, 2010.
- [3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", LSM '11 Proceedings of the workshop on Languages in Social Media, Columbia University New York, ISBN: 978-1-932432-96-1, pp. 30-38, 2011.
- [4] Y. Bae, H. Lee, "A Sentiment Analysis of Audiences on Twitter: Who Is the Positive or Negative Audience of Popular Twitterers?", In: Lee G., Howard D., Ślęzak D. (eds) Convergence and Hybrid Information Technology. ICHIT 2011. Lecture Notes in Computer Science, vol. 6935. Springer, Berlin, Heidelberg, 2011.
- [5] A. C. E. S. Lima and L. N. de Castro, "Automatic sentiment analysis of Twitter messages," 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), Sao Carlos, Brazil, pp. 52-57, 2012, doi: 10.1109/CASoN.2012.6412377.
- [6] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By, "Sentiment Analysis on Social Media," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey, pp. 919-926, 2012, doi: 10.1109/ASONAM.2012.164.
- [7] H. Kang, S. J. Yoo, D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, vol. 39(5), pp. 6000-6010, 2012, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2011.11.107>.
- [8] M. Ghiassi, J. Skinner, D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", in Expert System with Applications , ScienceDirect, vol. 30(16), pp. 6266-6282, Nov 2013.
- [9] A. Hassan, A. Abbasi, Daniel Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework", Social Computing (SocialCom), 2013 International Conference, IEEE, INSPEC Accession Number - 14024707, PP 357 – 364, 8-14 Sept. 2013.
- [10] E. Haddia, X. Liua, Y. Shib, "The Role of Text Pre-processing in Sentiment Analysis" , in the International Conference on Information Technology and Quantitative Management, Elsevier, Procedia Computer Science, Dec 2013.
- [11] G. Patil, V. Galande, V. Kekan, K. Dange, "Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, Issue 1, pp. 42–49, January 2014.
- [12] I. Dilrukshi, K. de Zoysa, "A Feature Selection Method for Twitter News Classification", in International Journal of Machine Learning and Computing, vol. 4, Issue. 4, August 2014, DOI: 10.7763/IJMLC.2014.V4.438, pp 365-370.
- [13] BAC Le, and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques", Advanced Computational Methods for Knowledge Engineering, Advances in Intelligent Systems and Computing 358, DOI: 10.1007/978-3-319-17996-4_25, Springer International Publishing Switzerland, pp. 279-289, 2015.
- [14] B. S. Dattu, Prof. Deipali V. Gore, "A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6 (6), pp 5358 -5362, 2015.
- [15] O. Kolchyna, Th'arsis T. P. Souza, Philip C. Treleaven and Tomaso Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination", Cornell University Library, 18 Sep 2015, arXiv :1507.00955.
- [16] J. Prusa, Tahhi M. Khoshgoftaar, David J. Dittman, "Using Ensemble Learners to Improve Classifier Performance on Tweet Sentiment Data", Information Reuse and Integration (IRI), IEEE International Conference, pp. 252–257, 13-15 Aug. 2015, INSPEC Accession Number: 15556647.
- [17] K. L. Devi, P. Subathra, P. N. Kumar, "Tweet Sentiment Classification Using an Ensemble of Machine Learning Supervised Classifiers Employing Statistical Feature Selection Methods", Proceedings of the Fifth International Conference on "Fuzzy and Neuro Computing (FANCCO - 2015)", vol. 415, pp. 1-13, Nov 2015.
- [18] Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis", in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, pp. 1318-1325, 2015.
- [19] Joseph D. Prusa, Taghi M. Khoshgoftaar, David J. Dittman, "Impact of Feature Selection Techniques for Tweet Sentiment Classification", Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, pp. 299-304, 2015.
- [20] R. Mansour, M. F. A. Hady, E. Hosam, H. Amr., A. Ashour, "Feature Selection for Twitter Sentiment Analysis: An Experimental Study", International Conference on Intelligent Text Processing and Computational Linguistics, Springer, pp. 92-103, 2015.
- [21] O. Abdelwahab, M. Bahgat, C. J. Lowrance and A. Elmaghraby, "Effect of training set size on SVM and Naive Bayes for Twitter sentiment analysis," 2015 IEEE International Symposium on Signal Processing

- and Information Technology (ISSPIT), Abu Dhabi, United Arab Emirates, pp. 46-51, 2015, doi: 10.1109/ISSPIT.2015.7394379.
- [22] S. Akter and M. T. Aziz, "Sentiment analysis on Facebook group using lexicon based approach", 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, pp. 1-4, 2016, doi: 10.1109/CEEICT.2016.7873080.
- [23] M. Bouazizi, T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter", IEEE Access 4, pp. 5477-5488, 2016.
- [24] P. Grandin, J. M. Adan, "Piegas: A systems for sentiment analysis of tweets in Portuguese", IEEE Latin America Transactions, vol. 14(7), pp. 3467-3473, 2016.
- [25] Nádia Félix Felipe da Silva, Luiz F.S. Coletta, Eduardo R. Hruschka, Estevam R. Hruschka, "Using unsupervised information to improve semisupervised tweet sentiment classification", Elsevier journal of Information Sciences, vol. 355-356, pp. 348-365, 10 August 2016.
- [26] A. Tripathy, Ankit Agrawal, Santanu Kumar Rath, "Classification of sentiment reviews using n-gram machine learning approach", in Expert System With Applications, ELSEVIER, pp. 117-126, 2016.
- [27] A. Krouska, C. Troussas, M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis", ResearchGate Conference, July 2016, DOI: 10.1109/IISA.2016.7785373.
- [28] A. Kashif, H. Dong, A. Bouguettaya, A. Erradi, R. Hadjidj, "Sentiment Analysis as a Service: A Social Media Based Sentiment Analysis Framework", in IEEE International Conference on Web Services (ICWS), Honolulu, HI, USA: IEEE, 2017.
- [29] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq S. Lee, "Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression", 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, pp. 138-140, 2017, doi: 10.1109/ICTC.2017.8190959.
- [30] M. R. Huq, Ahmad Ali, Anika Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM", In (IJACSA) International Journal of Advanced Computer Science and applications, vol. 8, Issue. 6, pp. 19-25, 2017.
- [31] M. Ahmad, Shabib Aftab, Iftikhar Ali, "Sentiment Analysis of Tweets using SVM", International Journal of Computer Applications (0975 - 8887), vol. 177, Issue 5, pp. 25-29, Nov 2017.
- [32] J. Brandon, J. Deng, "Sentiment Analysis of Tweets for the 2016 US Presidential Election", in IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA: IEEE, 2017.
- [33] R. Wijayanti, A. Arisal, "Ensemble approach for sentiment polarity analysis in user-generated Indonesian text", in 2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Jakarta, pp. 158-163, 2017.
- [34] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis", IEEE Access, Feb 2017.
- [35] M. A. Rahman, V. Mezhyuev, M. Z. A. Bhuiyan, S. N. Sadat, S. A. B. Zakaria, N. Refat, "Reliable decision making of accepting friend request on online social networks", IEEE Access 6, 9484-9491, 2018.
- [36] Z. Jianqiang, G. Xiaolin, Z. Xuejun, "Deep convolution neural networks for twitter sentiment analysis", IEEE Access 6, 23253-23260, 2018.
- [37] K. Tago, Q. Jin, "Influence analysis of emotional behaviors and user relationships based on twitter data", Tsinghua Science and Technology vol. 23(1), pp. 104-113, 2018.
- [38] I. Victoria, M. Sharmina, K. Malik, R. Batista-Navarro, "Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers", in Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS): IEEE, pp. 95- 98, 2018.
- [39] C. Troussas, A. Krouska and M. Virvou, "Evaluation of ensemble-based sentiment classifiers for Twitter data, in 7th International Conference on Information", Intelligence, Systems & Applications (IISA), Chalkidiki, pp. 1-6, 2016.
- [40] M. M. Fouad, T. F. Gharib, A. S. Mashat, "Ecient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble", in International conference on Advanced Machine Learning and Applications, vol. 723, pp. 517-527, Jan 2018.
- [41] Y. Emre Isik, Y. Görmez, O. Kaynar, Z. Aydin, "NSEM: Novel Stacked Ensemble Method for Sentiment Analysis", 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, pp. 1-4, 2018.
- [42] F. T. Giuntini, et al., "How Do I Feel? Identifying Emotional Expressions on Facebook Reactions Using Clustering Mechanism", in IEEE Access, vol. 7, pp. 53909-53921, 2019, doi: 10.1109/ACCESS.2019.2913136.
- [43] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression", in IEEE Access, vol. 7, pp. 163677-163685, 2019, doi: 10.1109/ACCESS.2019.2952127.
- [44] S. Vashishtha, S. Susan, "Fuzzy rule based unsupervised sentiment analysis from social media posts", Expert Systems with Applications, vol. 138, 2019, 112834, ISSN 0957-4174.
- [45] K. Elshakankery, M. F. Ahmed, "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis", Egyptian Informatics Journal, vol. 20, Issue 3, 2019, pp. 163-171, ISSN 1110-8665, <https://doi.org/10.1016/j.eij.2019.03.002>.
- [46] Martin-Domingo, Luis, Juan Carlos Martin, and Glen Mandsberg, "Social Media as a Resource for Sentiment Analysis of Airport Service Quality (ASQ).", Journal of Air Transport Management, 2019.
- [47] M. Naz, K. Zafar, A. Khan, "Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm", vol. 4, no. 2, pp. 1-13, May 2019. <https://doi.org/10.3390/data4020076>.
- [48] J. J. Bird, A. Ekárt, C. D. Buckingham, D. R. Faria, "High Resolution Sentiment Analysis by Ensemble Classification", In: Arai K., Bhatia R., Kapoor S. (eds) Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing, Springer, Cham, vol. 997, June 2019.
- [49] M. Khader, A. Awajan, and G. Al-Naymat, "The Impact of Natural Language Preprocessing on Big Data Sentiment Analysis", International Arab Journal of Information Technology, vol. 16, pp. 506-513, 2019.
- [50] R. Ahujaa, A. Chuga, S. Kohlia, S. Guptaa, and P. Ahujaa, "The Impact of Features Extraction on the Sentiment Analysis", International Conference on Pervasive Computing Advances and Applications - PerCAA 2019, Procedia Computer Science, vol. 152, pp. 341-348, 2019.
- [51] M. bibi, M. S. A. Nadeem, I. H. Khan, S. Shim, I. R. Khan, U. Naqvi and W. Aziz, "Class Association And Attribute Relevancy Based Imputation Algorithm To Reduce Twitter Data For Optimal Sentiment Analysis", Oct 2019, pp. 1365535- 1365544, vol. 7.
- [52] M. Bibi, W. Aziz, M. Almarashi, I. H. Khan, M. S. A. Nadeem and N. Habib, "A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis," in IEEE Access, vol. 8, pp. 68580-68592, 2020, doi: 10.1109/ACCESS.2020.2983859.
- [53] Z. Kermani, F., Sadeghi, F. & Eslami, E., "Solving the twitter sentiment analysis problem based on a machine learning-based approach.", Evol. Intel. 13, pp. 381-398, 2020. <https://doi.org/10.1007/s12065-019-00301-x>.
- [54] Esraa A. Afify, Ahmed Sharaf Eldin and Ayman E. Khedr, "Facebook Profile Credibility Detection using Machine and Deep Learning Techniques based on User's Sentiment Response on Status Message", International Journal of Advanced Computer Science and Applications(IJACSA), vol. 11(12), 2020.
- [55] S.R. George, P. Sujith Kumar, S.K. George, "Conceptual Framework Model for Opinion Mining for Brands in Facebook Engagements Using Machine Learning Tools.", In: Fong S., Dey N., Joshi A. (eds) ICT Analysis and Applications. Lecture Notes in Networks and Systems, vol. 154. Springer, Singapore, 2021.

Efficient Security Model for RDF Files Used in IoT Applications

Mohamed El kholy¹

Computer Engineering Department
Pharos University in Alexandria
Alexandria, Egypt

Abdel baes Mohamed²

Computer Engineering Department
AASTMT
Alexandria, Egypt

Abstract—The openness environment of IoT ecosystem arises several security and privacy issues. However, the huge amount of data produced by several IoT devices restricts using traditional security methods. Another security challenge for IoT system is the interoperability between heterogeneous IoT devices. Semantic Web has risen as a promising technology that provides semantic annotations allowing interoperability between IoT devices. Semantic web uses RDF triples to allow semantic data exchange between heterogeneous applications. Hence, RDF files used in IoT systems require specific security mechanism that regards large data size as well as rapidly data updates. The proposed work introduces a security novel that provides RDF files with a fine grained partial encryption. The proposed method allows applying security for the sensitive parts of RDF files without affecting the public parts. Encryption metadata is stored in a container related to each individual sensitive triple. Thus accessing public data in RDF file is not affected with the encryption overheads. A motivation scenario for privacy in a smart city is used to evaluate the proposed method. Experimental results showed that the proposed methodology enhances the access time of RDF triples from 10.4 msec to 6.2 msec. Moreover the proposed method facilitates integration of separated parts of a RDF graph together. The empirical evaluation proved the enhancement in efficiency and flexibility by applying the proposed method to RDF files used in IoT systems. Moreover the insensitive triples in RDF files are not affected with the security overheads.

Keywords—*Semantic Web; Internet of Things (IoT); resource description framework (RDF); smart cities; security mechanism; web ontology language (OWL); partial encryption; SPARQL protocol and RDF query language (SPARQL); data encryption standard (DES) component*

I. INTRODUCTION

IoT ecosystem connects several heterogeneous devices scattered all over the world [1]. IoT system relies on a set of sensors and actuators. Sensors are responsible of collecting data from the surrounding environment [2]. While actuators act on performing different actions for controlling devices [3]. Sensors and actuators are connected to the Internet by several heterogeneous protocols [1]. Each IoT device has its own hardware manufacture and can transmit and receive data with a specific defined format [4]. Interoperability between these heterogeneous devices at the hardware level or at the physical network layer is strongly complex [2, 3]. Moreover, different IoT devices use different protocols for data transfer [2]. The traditional protocols for ether net and Wi-Fi are used as well

as other protocols that maintain power saving such as zigbee and blue tooth [3]. Thus, interoperability between different IoT devices is considered one of the significant challenges for the success of IoT ecosystem.

On the other hand Semantic Web modified the Internet contents from documents for humans to read towards information for machines to manipulate [5]. Semantic web uses RDF triples as well as defined Web Ontology Language (OWL) to provide defined meaning to data [6]. RDF triples consists of subject, predicate, and object [5]. The subject identifies anything wanted to be described, while the predicate identifies the property or the attribute of description. The object is the value of the identified property. Representing data in RDF style allows different machines to get useful information about the status of the described subjects [7]. Thus heterogeneous machines can exchange data and also understand the meaning of such data [8]. Hence, Semantic RDF triples can provide significant benefit to IoT system [6]. RDF annotations allows IoT ecosystem to structure and enrich data coming from different IoT devices. Converting signals collected from IoT devices to RDF data allows applying semantic calculations on these signals [9]. Hence, using RDF for representing data provides IoT system with the required interoperability between different heterogeneous devices.

IoT system allows monitoring the surrounding environment and performing intelligent actions on behalf of human [10, 11]. Thus, IoT systems are widely used in smart cities. Such automated environment arise significant challenges of privacy and security. Several devices in smart homes submit different types of data. Among these data there exist private data that should not be available to public users [12]. Supplying IoT data with security and privacy is a key challenge for the success of smart cities. On the other hand, transferring IoT data to RDF triples increases the size of data [13]. Hence, traditional approaches of encryption and decryption are not suitable for RDF data used in IoT system. Traditional encryption and decryption techniques need high computational power resulting in high latency time [14]. Hence, such techniques are not suitable for IoT systems that are characterized with rapid data updates and the need to take quickly decisions [15].

The proposed work contributes in providing RDF files with a technique that allows partial fine grained encryption for RDF triples. The proposed technique links the encryption metadata to the encrypted triple directly without any

overheads to the main RDF file. Thus, unlike the traditional encryption containers, the insensitive data is not affected by the encryption overheads. The proposed technique has two significant enhancements; the first is shorter access time to insensitive data to RDF files even if it contains another encrypted data. The second is enabling to integrate the encrypted triples to another RDF file directly without processing the complex metadata in the RDF header. At the sender side the sensitive triples are encrypted and the encryption metadata is linked to each encrypted triple individually. At the recipient side the encrypted triples are decrypted using the schema send with each individual triple. Hence, a public user could access the public part of the encrypted RDF file without any encryption overheads. While the sensitive parts is restricted to authorized users only who can decrypt the sensitive triples. Hence, RDF triples are provided with the required security and privacy constrains while maintaining performance and efficiency aspects. The proposed work is limited to IoT sub systems that communicate by sending RDF files.

The remaining of the paper is organized as follows; a literature review is presented in section two. Section three introduces the problem definition. The proposed solution is discussed in section four. Finally the proposed model is evaluated in section five.

II. LITERATURE REVIEW

A significant number of published works discuss providing robust security for IoT systems without regarding the drawbacks of increasing data size and data access time. Other researches focus on securing only defined scenarios of using IoT systems. Securing RDF stores is also an attractive area for researchers. Several mechanisms for data encryption and access control are defined to provide security for semantic data. However, a little amount of work discusses securing semantic data associated with IoT systems. This literature review discusses the work done to secure RDF data in the spirit of openness and heterogeneous environment of IoT ecosystem.

Fernández et al. [16] defines a fine grained security for RDF triples. Their mechanism of encryption depends on the triples rather than a dedicated mediator. Their work combined symmetric and asymmetric encryption to reach high efficient security for RDF triples. They applied functional encryption to RDF data. The functional encryption allows the encrypted RDF triples to self-enforce its access restrictions. The Authors defined an encryption function derived from the RDF graph and randomly generated seeds. This function is used to construct a triple encryption vector for each RDF triple. Their encryption technique provides high security, however it is inefficient for large size of data. Moreover encrypting triples in such complicated technique makes security recovery challenging in case of different errors. Thus such security technique is not suitable for IoT environment.

Prajit Kumar Das et al. [17] designed a security framework that regard different security policies for data transfer between IoT devices. Authors represent security policies in semantically annotated statements. The framework defines different policies for access control depending on user attribute and the context

of IoT devices usage. Such policies are represented semantically using OWL to allow different computer machines to deal with it. Access to data associated with an IoT resource depends on the context of requesting this data. The context includes predefined relationship between user, and the RDF triple (subject, predicate and object). Thus access control is granted by particular permission for a specific user to use specific RDF at a specific situation. The framework shows high complexity and lacks flexibility needed for IoT open environment. The context of using IoT devices is a subject of continuous change. Thus defining access control according to the context will decrease the efficiency of using IoT data. Moreover it is complicated to include all scenarios of using IoT devices.

Pedro Gonzalez-Gil et al. introduces data-security ontology for IoT [18]. Authors represent a common vocabulary describing the practical security aspects related to data access that is relevant to producers and consumers. They defined two main classes one for secure data and the other for access control. The secure data is divided into hidden data and encrypted data, while the access control defines the authority for each party to access the secure data. Their work integrates security metadata such as access control and data protection to the traditional semantic data annotation. Then they used triples to describe the security aspects of different parts of data. Their work focused on defining the security requirements rather than applying these security to semantic data. Moreover, such mechanism is complicated to be implemented in a rapidly data changing environment of IoT ecosystem.

Another significant contribution for semantic security was done by Guangquan Xu et al. [19]. They defined a set of different Ontologies to describe security. Their work used Ontology to describe context and other Ontology to describe network attack as well as Ontology for system vulnerability. The network attack Ontology allows detecting complex attacks using a set of inference rules. While the vulnerabilities Ontology is responsible to detect elements exposed to danger to warn about this danger and its possible attacks. However, using different Ontologies increases the system complexity and increases the size of semantic data.

III. PROBLEMS DEFINITION AND MOTIVATION EXAMPLE

A. Problems Associated with Securing RDF Data used in IoT Systems

The proposed work contributes in filling the knowledge gap for the methodologies of securing RDF stores while maintaining its openness and semantic features. Traditional security approaches are not suitable to provide RDF triples with the required security and data privacy while maintaining openness features. RDF stores provide IoT system with semantic meaning that allows linking data from different IoT devices [20, 21]. In such an open and heterogeneous environment that lacks human monitoring security and privacy requirements increase significantly [22]. Methods for specifying the role of each agent to access or to use specific pieces of data are considered a significant challenge.

1) *First problem:* RDF stores lack a trustworthy infrastructure for specifying access control. Such problem is a

reflect of the openness environment of semantic RDF triples. Another challenge will appear even a robust access control mechanism is applied to RDF stores. RDF data is transferred in non-secure channels so it is not safe from different sniffing attacks [23]. Thus using access control to secure RDF stores cannot provide the required security and privacy for IoT system.

2) *Second problem:* Traditional RDF encryption solutions are not suitable to the openness and rapidly changing environment of IoT system. IoT includes billions of sensors that transmit huge volume of data which is frequently changed [24]. Encryption and decryption of such a huge amount of data consume high computational power and significantly affect performance. Moreover, such approach affects real time applications that depend on the speed of reasoning IoT data. RDF partial encryption is used to solve such a problem but still include significant drawbacks. A significant number of RDF files is serialized in XML files to allow interaction between heterogeneous machines [25]. RDF partial encryption selects the sensitive data and encrypts it, and stores the encryption metadata in XML file header. Thus increase the size of RDF files when represented in XML files and needs more time to process data in the file headers. Moreover IoT environment is characterized with rapidly data updates which will needs continuous updates to the file header consuming more computational power. Another drawback in traditional partial encryption approach is that IoT data includes a small size of sensitive data and a large size of public data. In traditional partial encryption approach even accessing public data is affected with processing the encryption metadata which is stored in the XML file header.

B. Motivation Example

Smart cities include huge number of sensors that are responsible of monitoring the surrounding environment. These sensors vary from temperature sensors, humidity sensors, cameras and others. Sensors are distributed everywhere on the streets, over the buildings and also inside homes and homes' gardens. Cameras on home gardens record different images from different angles and may be supported with facial recognition facilities. Data emitted by cameras include private data for the home owner. However, for security reasons police station or city authority should be able to access such data in specific intervals of time. The data associated with the home owner also includes sensitive parts such as the bank account or bank balance. To be more general the motivation example is based on sensor that emits data that should be accessed by specific users and restricted from another. Fig. 1 represents a general overview of sensors used in IoT systems. Thus as mentioned before the size of private data is small compared with the total size of monitoring data. Applying security features for the whole RDF file is not efficient. Instead, security should be applied only to the sensitive part of data.

IV. PROPOSED RDF PARTIAL ENCRYPTION MODEL FOR IoT SYSTEM

The proposed model perform a fine grained partial encryption for RDF files in which sensitive triples are encrypted while non-sensitive triples are represented in plain text. Unlike traditional approaches, encryption metadata is linked to individual RDF triples not to the whole graph. Thus each encrypted triple encapsulates its cipher text as well as its encryption metadata. RDF documents contain a small part of sensitive data and a large part of public data. Hence, it is more efficient to apply security features to the sensitive triples and store the encryption metadata as extension of the encrypted triple. Any access to public data is not affected by the metadata of the encrypted triples. Thus the openness and rapidly changing environment of IoT is not affected. Moreover storing encryption metadata as separated triples extended from the encrypted triple increase coherence features of RDF graphs. RDF graphs can be divided to smaller sub graphs. Then sub graphs can be shared in another RDF graph.

A. First Step: Selecting Private Data

The In the first step all the sensitive data are selected over the RDF graph. To illustrate the idea the motivation scenario discussed in Section 3.2 is used. As shown in Fig. 1 the RDF graph presents an image captured by a fixed camera installed in a home garden to monitor the motion of persons. The captured images include private data that should be limited to the home owners. However at specific time the police station may require to use these images. Analyzing the RDF graph shows that sensitive data is limited to the image URL which include the link to the captured image. Other data does not include high level of privacy such as camera type or file format. Thus the first encrypted triple is the object in the following triple.

(Subject: captured image, Predicate: saved in, Object: image URL)

Another sensitive data related to the person how owns the home is his bank account. The second encrypted triple is the subject and object in the triple (Subject: bank account, Predicate: has balance, Object: balance).

Thus, the output of the first step determents which fragments of data should be encrypted. These fragments are named Sensitive Triples (ST) while the remaining triples are named Plain Triples (PT). The proposed work does not restrict a method to select sensitive data. ST selection may be done statically by enumerating the encryption fragment in the document before run time. Selection also can be done dynamically during run time by specifying selection patterns which check specific properties.

B. Second Step: Encrypting the Selected Fragment

After defining ST and PT an Encryption Function (EF) is added to each RDF ST in the RDF file. Thus for a graph G of RDF triples, $ST \in G$ is an encrypted triple $EF(ST) = (ST, \text{has encryption container}, E_c)$ where ST is the sensitive triple that will undergo encryption, as a whole triple or parts of it. E_c is the encryption container associated with the sensitive triple.

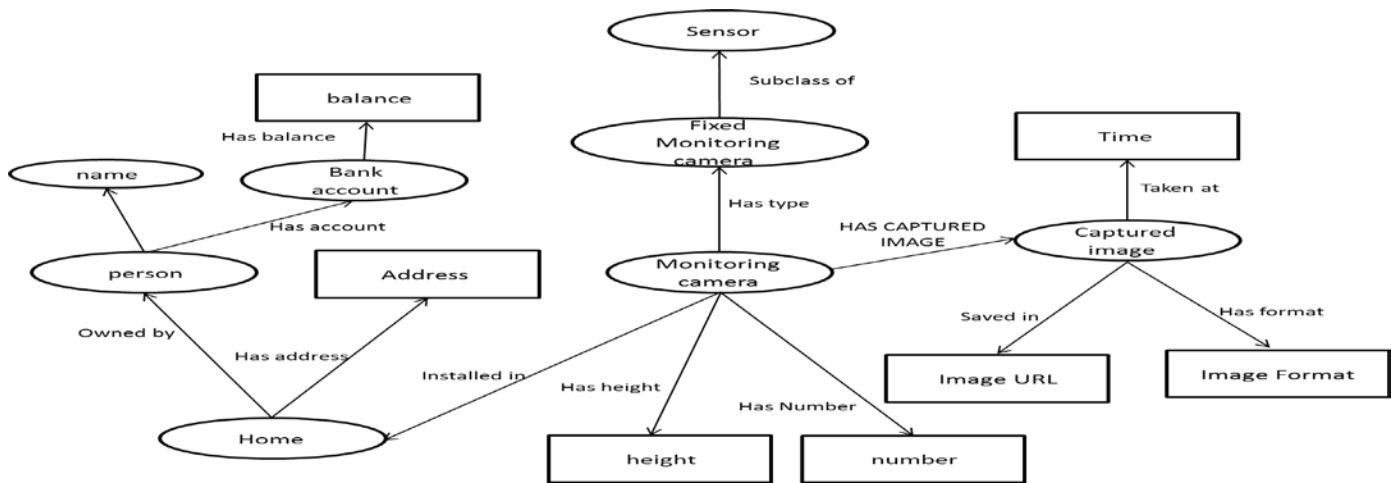


Fig. 1. RDF Representation of IoT Sensors in a Smart City.

The encryption container contains all the required encryption metadata. The encryption container contains a number specifying which part of the triple to be encrypted, as well as the encryption key and the encryption method. It also contains reference to the triple subject, predicate, and object as plain or cipher text according to the encryption number E_n . The encryption number has a value from 1 to 7 as there are only seven possible probabilities to choose the EF from each triple (Subject, Predicate, Object) (S, P, O). $E_n = 1$ for O only, 2 for P only 3 for O,P, 4 for S only, 5 for S, O, 6 for S,P, 7 for S, P, O. The encryption key K_t is associated with the encrypted triple and is encrypted with the public key of the authorized user.

To clarify the parameters of the encryption function, it is applied to the motivation example in section 3.2 as follows:

(Subject: captured image, Predicate: saved in, Object: image URL)

The sensitive data is the URL of the image which is the object so the encryption function is defined as follows:

EF (ST) = (Sensitive Triple, has container, E_c) then the encryption container has the following related properties:

(E_c , has number, E_n) E_n is the encryption number which has the value of 1 (only the object will be encrypted).

(E_c , has key, E_k) a new E_k is generated for each sensitive triple and is encrypted by the public key of authorized user.

(E_c , has method, DES) the object is encrypted with DES.

In this case the object will be in cipher text, while the subject and predicate will be in its plain text.

For the second sensitive data in the motivation example:

(Subject: bank account, Predicate: has balance, Object: balance)

The sensitive data are the bank account and the balance, which are the subject and object so the encryption function will be:

EF (ST) = (Sensitive Triple, has container, E_c)

(E_c , has number, E_n) E_n is the encryption number which has the value of 5 (Subject, Object)

(E_c , has key, E_k) a new E_k is generated for each sensitive triple.

(E_c , has method, DES) the subject and object is encrypted with DES

Thus each encrypted triple has an associated metadata which is inserted to the RDF graph as an encryption container. The encryption container contains a set of triples that presents the metadata of the original encrypted triple.

The encryption container includes three parts of metadata associated with the encrypted triple as well as the three triple subject, predicate, and object. First part is the encryption number which defines which parts of the triple are encrypted. Second part is the encryption key that used to perform symmetric encryption of the triple. Third part is the cipher text of the encrypted part of the triple. There is no need to encrypt all the data in the triple the sensitive parts are encrypted while public parts are presented in plain text.

C. Third Step: Decryption

Each reference to a sensitive triple will be directed to the encryption container associated with this triple Fig. 2. The decryption is done according to the parameters specified in the encryption container. Authorized recipient will perform asymmetric encryption for the triple key using their asymmetric private key. Then the triple session key is used to decrypt the cipher text of the triple. If a receiver does not have an appropriate triple key, the decryption fails. Public users who access public triples are not affected with the encryption process as the encryption containers are associated with sensitive triples only. Fig. 3 represents the decryption process.

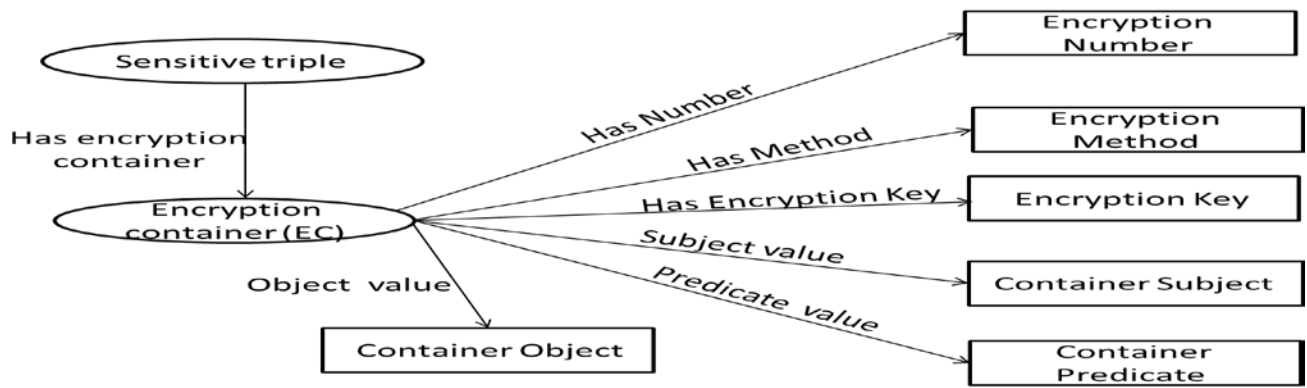


Fig. 2. Encryption Container.

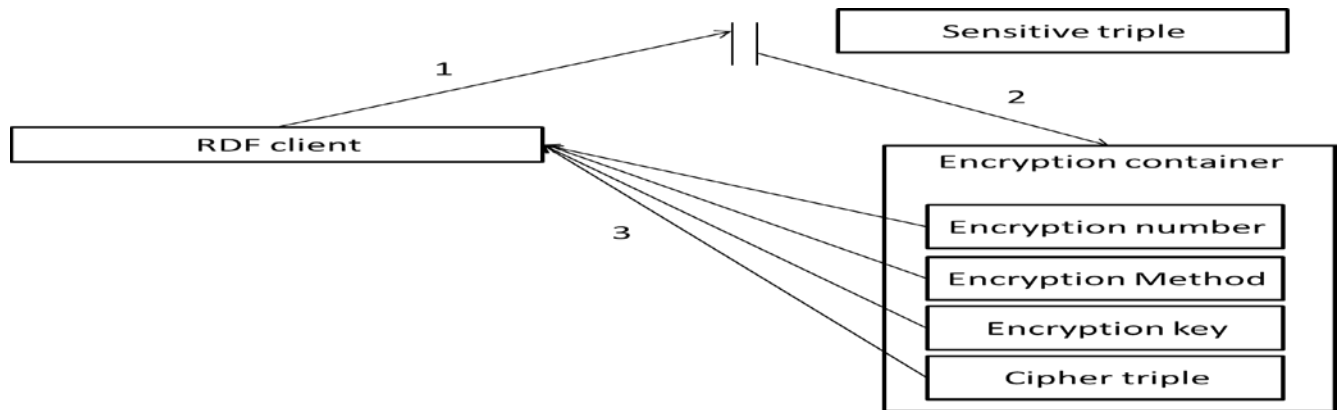


Fig. 3. Decryption Process of Encrypted RDF Triple.

V. IMPLEMENTATION AND EMPIRICAL EVALUATION

The aim of the evaluation is to proof the efficiency of the proposed method while maintaining strong security aspects for sensitive data. The motivation example of smart city is implemented using XML serialization. A data set including 1000 RDF triples is used to perform different calculations. The evaluation metrics are then compared with traditional partial RDF approach to proof the efficiency enhancement of the proposed work.

A. Evaluation Metrics

As the proposed RDF partial encryption method is mainly concerned with IoT echo system, it is supposed to maintain openness and rapidly data updates environment. Thus the encrypted RDF triples should allow rapid access time and rapid data updates. Other significant attributes in IoT systems is the ability to integrate several parts from different RDF graphs together. Hence, such attribute is considered while choosing evaluation metrics of the proposed work. Thus two metrics were chosen, the first is the access time, the second is the ability to integrate the RDF file with another files.

B. Implementation

To clarify the proposed method the motivation example was implemented using XML serialization. All the encryption metadata for each triple is linked to the triple itself rather than to the RDF header. Regarding the motivation example the object in the following triple contain private data.

Captured image: (hasimagefile) URL: image URL

According to the proposed work any access to the object of this triple will be directed to the Encryption container with the following properties:

Number = 1; as the encrypted part is the object

Triple key = K_t which is the session key encrypted by public key of authorized user.

Encryption method: DES

Cipher subject= the same as plain subject

Cipher predicate= the same as plain predicate

Cipher object= encrypted URL of the image

Thus the XML serialization of the sensitive triple is as shown below in listing 1.

```
</rdf:Description>
</rdf:Description
rdf:about="http://www.smartcity/.../CapturedImage/EncryptionContainer#">
<EC: Number> 1 </EC: Number>
<EC: tripleKey> EncryptionKey </EC: Number>
<EC: EncryptionMethod> DES </EC: EncryptionMethod>
<EC: PlanSubject>http://www.smartcity/.../CapturedImage#/IMG2013
</EC: planSubject>
<EC: PlanPredicate> Ci </EC: PlanPrdicate>
<EC: CipherObject> ##### </EC: CipherObject>
```

Listing. 1. XML Serialization of the Encryption Container.

C. Experimental Results

To evaluate the proposed method, 1000 triples of the proposed RDF graph of the motivation example were

implemented using Apache Jena API in Java. Then the RDF file was encrypted twice. First time, the RDF file was encrypted using the proposed method by applying the encryption metadata to individual sensitive triples. The other time the traditional encryption approach was applied to the RDF file by encrypting each triple and inserting the encryption metadata in the RDF file header. Then SPARQL query was used to access different sensitive and non-sensitive triple for the two RDF files. SPARQL query was applied to each triple in the file once, twice, and three times. The time to

get data was calculated for the two files. The results were clarified in Table I. Results prove the enhancement of access time for the proposed partial encryption method. Using the proposed approach any access to non-sensitive data in RDF file will not process the encryption metadata. As the metadata is encapsulated in encryption container associated with each encrypted triples only. However, traditional RDF encryption approach requires processing the header for each access to the file whether the data is sensitive or not.

TABLE I. ENHANCEMENT OF ACCESS TIME USING RDF THE PROPOSED APPROACH

Number of time of applying SPARQL queries to each triple in the file	Response time for proposed method	Response time for traditional RDF encryption
One time	2.5 msec	5.3 msec
Two times	3.8 msec	8.2 msec
Three times	6.2 msec	10.4 msec

D. Discussion

To illustrate the efficiency of the proposed model it was compared with existing work. The results in Table I compare between the proposed work and traditional state-of-the-art methods for RDF encryption methods. It is observed that the proposed technique decreases the response time significantly for more than 50%. Thus the proposed technique provides IoT application with high respond time for SPARQL queries while maintain high level of security and data privacy. Moreover, the capsulation of encryption metadata in triples associated with the encrypted triple allows linking this encrypted triple to another RDF graph Moreover the proposed technique allows flexible integration of RDF triples from one RDF file to another without consuming time in processing metadata in RDF header.

VI. CONCLUSION AND FUTURE WORK

The proposed work provides RDF files with a fine grained partial encryption method suitable to be used in IoT ecosystem. The proposed method allows applying security aspects for each RDF triple individual reducing the time to process encryption metadata while accessing non-sensitive triples. The benefit of such approach increases in IoT systems hence RDF files used in IoT system include small size of sensitive data and large size of public data. Thus the proposed method maintains the security of sensitive triples while enhancing the access time of public data.

Moreover the proposed method provides IoT systems with the ability to integrate triples from different RDF file together to deal with different environments. The proposed approach supports such requirement as the security metadata is related to each individual triple rather than to the file as a whole. A future work will analyze the ability to compress XML files that includes encrypted RDF triples. Our future work will also discuss the ability to select the sensitive data dynamically at run time

REFERENCES

[1] Sankar Mukherjee, G.P. Biswas, "Networking for IoT and applications using existing communication technology," Egyptian Informatics Journal, Volume 19, Issue 2, 2018, Pages 107-127, ISSN 1110-8665, <https://doi.org/10.1016/j.eij.2017.11.002>.

[2] Michael Haslgrübler, Peter Fritz, Benedikt Gollan and Alois Ferscha, "Getting through: modality selection in a multi-sensor-actuator industrial IoT environment" IoT '17: Proceedings of the Seventh International Conference on the Internet of Things October 2017 Article No.: 21 Pages 1–8 <https://doi.org/10.1145/3131542.3131561>.

[3] Ramadevi Chappala, Ch.Anuradha and P. Sri Ram Chandra Murthy, "Adaptive Congestion Window Algorithm for the Internet of Things Enabled Networks" International Journal of Advanced Computer Science and Applications(IJACSA), 12(2), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0120214>.

[4] Minhaj AhmadKhan, KhaledSalah, "IoT security: Review, blockchain solutions, and open challenges" Future Generation Computer Systems, May 2018, Pages 395-41 <https://doi.org/10.1016/j.future.2017.11.022>.

[5] Sabrina Kirrane, Serena Villata, Mathieu d'Aquin, Mathieu d'Aquin, Sabrina Kirrane, Serena Villata, "Privacy security and policies: A review of problems and solutions with semantic web technologies", Semantic Web, vol. 9, pp. 153, 2018.

[6] Haytham Al-Feel, Hanaa Ghareib Hendi and Heba Elbeh, "Enrichment Ontology with Updated user Data for Accurate Semantic Annotation" International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0101223>.

[7] N. Seydoux, K. Drira, N. Hernandez, T. Monteil, "Capturing the Contributions of the Semantic Web to the IoT: A Unifying Vision," Semantic Web Technologies for the Internet of Things Workshop colocated with 16th ISWC-2017, (2017).

[8] Kirrane, S., Mileo, A., Decker, S, "Access control and the resource description framework: a survey," SemanWeb8(2), 311–352 (2017). doi:10.3233/SW 160236. <http://dx.doi.org/10.3233/SW-160236>.

[9] Uceda-Sosa R, Srivastava B, Schloss RJ, "Building a highly consumable semantic model for smarter cities," In: Proceedings of the AI for an Intelligent Planet on - AIIP '11. New York, New York, USA: ACM Press; 2011:1-8. doi:10.1145/2018316.2018319.

[10] Mamdough Alenezi, Khaled Almustafa, Khalim Amjad Meerja, "Cloud based SDN and NFV architectures for IoT infrastructure," Egyptian Informatics Journal, Volume 20, Issue 1, 2019, Pages 1-10, ISSN 1110-8665, <https://doi.org/10.1016/j.eij.2018.03.004>.

[11] Brambilla, M., Umuhoza, E. & Acerbis R, "Model-driven development of user interfaces for IoT systems via domain-specific components and patterns," Journal of Internet Services and Applications 8, 14 (2017). <https://doi.org/10.1186/s13174-017-0064-1>.

[12] A. Gharaibeh et al., "Smart Cities: A Survey on Data Management, Security, and Enabling Technologies," in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2456-2501, Fourthquarter 2017, doi: 10.1109/COMST.2017.2736886.

[13] Bhavani Thuraisingham, "Security standards for the semantic web," Computer Standards & Interfaces Volume 27, Issue 3, March 2005, Pages 257-268 <https://doi.org/10.1016/j.csi.2004.07.002>.

- [14] B. Lalitha and G. Murali, "Implementing deduplication technique for RDF files with enhanced security using multi cloud servers," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 3618-3621, doi: 10.1109/ICECDS.2017.8390137.
- [15] A. Esfahani et al., "A Lightweight Authentication Mechanism for M2M Communications in Industrial IoT Environment," in IEEE Internet of Things Journal, vol. 6, no. 1, pp. 288-296, Feb. 2019, doi: 10.1109/JIOT.2017.2737630.
- [16] .Kaleem Razzaq Malik, Yacine Sam, Majid Hussain, Abdelrahman Abuarqoub, "A methodology for real-time data sustainability in smart city: Towards inferencing and analytics for big-data, Sustainable Cities and Society," Volume 39, 2018, Pages 548-556, ISSN 2210-6707, <https://doi.org/10.1016/j.scs.2017.11.031>.
- [17] P. K. Das, S. Narayanan, N. K. Sharma, A. Joshi, K. Joshi and T. Finin, "Context-Sensitive Policy Based Security in Internet of Things," 2016 IEEE International Conference on Smart Computing (SMARTCOMP), St. Louis, MO, 2016, pp. 1-6, doi: 10.1109/SMARTCOMP.2016.7501684.
- [18] N. Yorino, A. Muhammad, Y. Sasaki, Y. Zoka, "Robust Power System Security Assessment under Uncertainties Using Bi-Level Optimization," IEEE Trans. on Power Syst., Vol. 33, No. 1, pp. 352-362, Jan. 2018.
- [19] G. Xu, Y. Cao, Y. Ren, X. Li and Z. Feng, "Network Security Situation Awareness Based on Semantic Ontology and User-Defined Rules for Internet of Things," in IEEE Access, vol. 5, pp. 21046-21056, 2017, doi: 10.1109/ACCESS.2017.2734681.
- [20] Vogt L., Baum R., Köhler C., Meid S., Quast B., Grobe P, "Using Semantic Programming for Developing a Web Content Management System for Semantic Phenotype Data," In: Auer S., Vidal ME. (eds) Data Integration in the Life Sciences. DILS 2018. Lecture Notes in Computer Science, vol 11371. Springer, Cham.
- [21] S. Benbernou, X. Huang and M. Ouziri, "Semantic-based and Entity-Resolution Fusion to Enhance Quality of Big RDF Data," in IEEE Transactions on Big Data, doi: 10.1109/TBDATA.2017.2710346.
- [22] Antonio Celesti, Maria Fazio, "A framework for real time end to end monitoring and big data oriented management of smart environments," Journal of Parallel and Distributed Computing," Volume 132,2019,Pages 262-273,ISSN 0743-7315, <https://doi.org/10.1016/j.jpdc.2018.10.015>.
- [23] Farhan Ullah, Muhammad Asif Habib, Muhammad Farhan, Shehzad Khalid, Mehr Yahya Durrani, Sohail Jabbar, "Semantic interoperability for big-data in heterogeneous IoT infrastructure for healthcare, Sustainable Cities and Society, "Volume 34,2017,Pages 90-96, ISSN 2210-6707, <https://doi.org/10.1016/j.scs.2017.06.010>.
- [24] H. Cai, B. Xu, L. Jiang and A. V. Vasilakos, "IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges," in IEEE Internet of Things Journal, vol. 4, no. 1, pp. 75-87, Feb. 2017, doi: 10.1109/JIOT.2016.2619369.
- [25] Joe Tekli, Nathalie Charbel, Richard Chbeir, "Building semantic trees from XML documents," Journal of Web Semantics, Volumes 37–38, 2016, Pages 1-24, ISSN 1570-8268, <https://doi.org/10.1016/j.websem.2016.03.002>.

Predictive Analysis of Ransomware Attacks using Context-aware AI in IoT Systems

Vytarani Mathane¹, P.V. Lakshmi²

Department of Computer Science and Engineering
GITAM University, Vishakhapatnam, India

Abstract—Ransomware attacks are emerging as a major source of malware intrusion in recent times. While so far ransomware has affected general-purpose adequately resourceful computing systems, there is a visible shift towards low-cost Internet of Things systems which tend to manage critical endpoints in industrial systems. Many ransomware prediction techniques are proposed but there is a need for more suitable ransomware prediction techniques for constrained heterogeneous IoT systems. Using attack context information profiles reduces the use of resources required by resource-constrained IoT systems. This paper presents a context-aware ransomware prediction technique that uses context ontology for extracting information features (connection requests, software updates, etc.) and Artificial Intelligence, Machine Learning algorithms for predicting ransomware. The proposed techniques focus and rely on early prediction and detection of ransomware penetration attempts to resource-constrained IoT systems. There is an increase of 60 % of reduction in time taken when using context-aware dataset over the non-context aware data.

Keywords—Ransomware; IoT; context-aware; machine learning; ontology

I. INTRODUCTION

IoT systems are distinct from others in that they are ubiquitous, heterogeneous in capabilities, and usually out in adversarial environments [1]. They are present in Industries, Medical centers, Smart cars, Smart homes, Smart cities, and supply chains [2]. Such IoT systems could be susceptible to multiple categories of attacks like Denial of Service (DoS), botnets, man in the middle, identity and data theft attacks, ransomware attacks given the less than the secured or controlled environment of deployment and quite often limited security capabilities [3]. Among all those ransomware attacks could be more impacting owing to attack methodology where victim systems become unusable until a ransom is paid, typically have attacker-defined timelines to respond, and can cause more monetary loss.

Ransomware attacks, one of the malware attacks affect all types of security issues availability which causes monetary losses, and sensitive information loss [4]. Crypto ransomware, locker ransomware, and hybrid ransomware are common types of ransomware [5][6]. In crypto-ransomware attacks, data files are encrypted and the decryption key is provided only after paying the ransom. In locker ransomware attacks, the resources are blocked and are released only after paying the ransom. In hybrid ransomware attacks, both concepts of crypto-ransomware and locker ransomware are used. BadRabbit, Petya-Escape, Scareware, Screen Lockers, WannaCry are some

famous ransomware attacks. By using Botnets, Social engineering, and malvertisement (malicious advertising) ransomware can penetrate IoT devices [6].

Context information of a typical device includes individuality, activity, location, time, and relation [7]. The prediction model has to utilize one or more of these categories to predict a ransomware attack. Location for tracking target and source, time for identifying the time of events occurring on the device, activity to find the set of events that leads to suspicious activity, relation to identifying the dependency between events, and individuality to identify the device through unique characteristics. These features are modeled and used for attack prediction. The context-aware prediction models can use different techniques such as graphs, anomaly detection, classification, clustering, etc.

II. RELATED WORK

AI algorithms are used for cyber defense, malware prevention, and advanced threat detection or prevention [8]. Machine learning is used to learn about the attacks and predict them or machine learning for learning attacks and pattern matching for predicting them [9]. MIT labs developed an AI2 platform to predict cyber-attacks using AI [10]. IoT systems use AI algorithms for attack and anomaly detection [11]. Support Vector Machine (SVM) model as it is good for predicting very specific attacks [9]. According to [12] they use SVM to detect and predict ransomware attacks. SVM is good for detecting zero-day attacks which are unknown [7][13].

Ransomware in IoT can affect the integrity, confidentiality, and availability of the system and can cause monetary losses and loss of sensitive information [14]. In [15] 18 families of ransomware are studied and developed a model for categorizing behavioral characteristics, which can be used to improve the detection of ransomware attacks. [16] uses weighted KNN machine learning technique to detect and predict ransomware attacks on software-defined networking. The author in [17] uses neural networks for detection of ransomware in Industrial IoT where there is a huge risk.

Context-awareness is achieved by [18] using Context ontologies and Ontology description logic to get dynamic context attributes. The author in [19] use known attack context profiles to detect specific attacks that are relevant to a particular context and to avoid false-positive alerts. Known attack context profiles are created using conditional entropy [20][21]. The author in [22] uses sensor ontologies according to the semantic needs of IoT solutions. Ontologies can be

categorized into device ontology, domain ontology, and estimation ontology. Semantic metadata like context, description of the sensor, and its configuration provides contextual information. The proposed paper uses a classification model using contextual features. Section 2 describes the related work on context-aware ransomware attack prediction. Section 3 describes the framework and design of predicting ransomware attacks. Section 4 describes the implementation and Section 5 shows the comparison of solutions using with and without context-aware features.

A. Anatomy of a Typical Ransomware Attack on IoT

There has been a significant amount of research on ransomware threats to the IoT segment [14][15][16][17] and it offers very significant insights into ransomware penetration in the area of IoT, attack vectors, methodologies, and few specific implementation details (like Windows APIs) used for attacks. From the analysis of previous ransomware attacks on IoT, the ransomware executes in the following stages:

- Stealth mode where ransomware attacker benefits as long as the attacked system does not detect ransomware.
- Suspicious mode where ransomware starts collecting vital stats required to assess the suitability of specific targets within the system and starts encrypting or locking those.
- Obvious mode where attacker and ransomware display messages to the victim with a chosen mechanism to the victim.

Predominately Windows-based workstations used in IoT grids, Proof of Concept on low-end IoT devices (smart bulbs, smart TV, etc.) are the typical IoT systems that are being attacked.

Crypto, Locker, and Hybrid are different types of ransomware attacks [6]. The attack vectors used are Content distribution, Social engineering, Malvertisement, Downloaders & botnets, Email phishing, and R-a-a-S (Ransomware as a service) [6].

The typical flow of successful ransomware attacks shows certain patterns. Attack made leveraging social engineering goes through a sequence where the victim is made to download ransomware, elevate privileges of ransomware and/or current user, exploit elevated privileges and & locally downloaded ransomware, establish a connection back to command center to make victim submit to demands. Another case of a ransomware attack on network interfaces goes through cyber scanning, enumeration, intrusion attempt, the elevation of privilege, perform malicious tasks, deploy malware/backdoor, delete forensic evidence, and exit.

Other patterns emerging out of existing data are also pointing to increased integrated fingerprinting as a part of mounting ransomware attacks. Such fingerprinting is used to vital data to decide on the usefulness of the content in extortion schemes, Usefulness of content is seen to be analyzed based on a multitude of factors like date & times of content creation, usage of content, location of content in the system, geolocation data, file extensions, file names & entropy of the content.

Since a significant number of attacks were targeted towards Windows OS-based IoT endpoints & IoT servers, there are few studies that leveraged analysis of the Windows APIs used and traversed during attacks to build prediction capabilities. So far most promising models are unfortunately based on very high-level sequences & context, e.g., specific sequences followed by ransomware attacks on network stack can hold the potential key to discovering IoT attacks in real-time.

B. Observations and Deductions from Past Studies

One can make three observations based on the analysis and outcomes of previous studies focused on ransomware attacks on IoT as below [14][15]:

- This anatomy of a typical ransomware attack on an IoT system as described in the prior section allows us to make a safe conclusion that it applies to a very specific section of IoT devices using Windows OS and hence use moderately powered CPUs and other resources. Ransomware attack prediction models built using such data are also applicable largely to such Windows-powered IoT systems.
- Content & hence content analysis plays an important role in the current attack landscape to detect victim system's suitability for exploitation followed by most suitable contents (files, directories, etc.) to execute one of ransomware attack technique (encryption, locking, hybrid).
- Third-social engineering plays a very significant role as an enabler to fetch ransomware into the victim system. The use of social engineering needs to factor in a user being present on the system to intentionally or unintentionally allows download and installation of such malware content. Without such a user being present, the ease of ransomware finding its way to the victim system reduces greatly.

Contrasting these observations of a type of IoT systems attacked, capabilities such systems possess, and attack vectors used with a low-end microcontroller and microprocessor-based IoT provides us a path forward. Such lower-end IoT systems could be a very interesting target because of several reasons:

- These systems could provide a much greater period of stealth and suspicious modes as those typically are unsupervised or do not have a human operator.
- These systems do control vital and critical nodes, operations within a grid and hard to pinpoint for fault analysis given the nature of deployment.
- These systems have tremendous heterogeneity lacking standard & widely used OS capabilities, underlying hardware, need to fine-tune ransomware for each such target system effectively making large scale deployment hard prospect.
- These systems typically do not store data but rather used as control endpoint or sensor endpoints, so the crypto category of ransomware attack does not have meaningful gains.

- These systems also provide smaller attack surface because of limited or none endpoint level user involvement, social engineering vulnerabilities.

Nonetheless these studies, patterns observed, APIs leveraged by ransomware can still be used to make some progress towards ransomware prediction capabilities for lower-end resource constrained IoT devices. Such devices are expected to be spread in a grid, typically control key elements in a grid and if attacked can also bring down large industrial critical infrastructures. It is only to be expected that attackers would want to leverage ransomware to cripple such ground-level IoT devices to maximize damage inflicted to scare victims into paying a ransom.

Segmentation is OSES used in a variety of low-end IoT systems; varying hardware also does not help much ability to build a predictive model as it leads to segmentation of data observed from such systems. One way is to up-level predictive models from specific APIs and capabilities to allow such heterogeneity of implementations.

III. DESIGN METHODOLOGY

Building on the previous section, we aim to provide a solution for predicting ransomware attacks in lower-end IoT systems (which has been largely neglected so far) using context-aware AI algorithms methodologies.

A. Building Context Parameters

Context is further defined considering the following factors: target IoT systems, deployment vectors, and attack vectors. Common and specific use cases for target IoT systems are sensor nodes, controllers to a specific function in the power grid, valve controllers on the dam, etc. All such use cases imply that to target specific capability, ransomware needs to collect information about ports, memory addresses, etc.

More prevalent methods of ransomware deployment like social engineering, malvertisement, email phishing does not make any sense to IoT, whereas the following methods can be leveraged to deploy ransomware to lower-end IoT systems: content distribution, downloaders & botnets, and Ransomware as a service (R-a-a-S). Attack Vectors consists of various events, activities, or APIs associated with the above deployment vectors including software, firmware update capabilities (system-specific APIs downloading new firmware packages and overwrite existing memory contents), connection requests /traffic in and out of the system on available transmission protocols (Wi-Fi, BT, etc.), port scanning (scanning for memory input/output port addresses) and use of cryptographic APIs & underlying accelerators or software libraries (typically OpenSSL AES encryption APIs).

B. Scope and Design of the Solution

Further narrowing on the scope of this proposed solution, attacker's entry attempts into an IoT system via TCP/IP or BT-like protocols is focused. This is in line with the strategy that prevention is better than cure and in as early stage as possible. Port scanning, scanning for cryptographic APIs indicate the attacker is already in the system and for the current discussion it is beyond the scope.

Context-awareness is achieved using context ontology and developing attack profiles. The data is provided to the AI models and the attack is predicted. The design of the solution includes the following main components: data collection, Context ontology (for feature extraction), attack context filters, Classification algorithm (for pre-diction), Result alert. Fig. 1 shows the design components of the proposed solution.

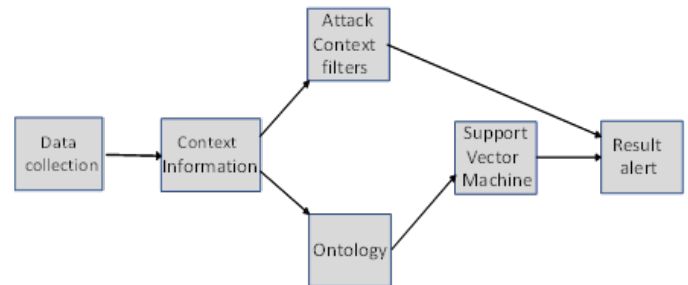


Fig. 1. Design Components of the Solution.

1) *Data collection*: IoT devices communicate via network stack, packets are collected using a network tool. Thus, collected data is used as training data which comprises benign and ransomware attack traffic. The data set collected from our testbed is represented in a JSON file format.

2) *Context ontology*: The collected data is represented using context ontology which follows logic and structure and automates the information retrieval. The context ontology uses the data collected from the data collection unit.

As mentioned above only a subset of known ransomware attack vectors (content distribution, downloaders & botnets, and Ransomware as a service) are likely to be used in attacking a typical industrial IoT. In this study, a context ontology for specific attack vectors of downloading ransomware or ransomware infected software images to IoT devices within the network is designed. The activity context information with features such as attacker, target IoT systems, and network events or activity towards downloading ransomware to target IoT device is built. Similarly, one could develop a context for using a content distribution like device configuration or parameters and Ransomware-as-a-service but those are out of scope for the present study.

3) *Attack filters*: The activity context information is used to create attack context profiles for classifier algorithms. The attack profiles are a set of features that are important to detect the attack. The feature selection is based on a set of rules followed to ensure detecting the attack. The feature vector can be represented using the equation (1):

$$F = \{f_1, f_2, \dots, f_n\} \quad (1)$$

Following Fig. 2 illustrates how to build attack filters for typical ransomware penetration of an IoT device within a specific IoT network using a download attack vector. Such a scenario comprises an attacker node attempting to impersonate authorized software or content distribution entity which would further attempt to detect possible target IoT devices by doing ipsweep and port scan for devices listening for software or content updates. Once such IoT systems are found, an attacker

node would attempt spoofing as a legitimate content provider and subsequently compromise IoT devices or devices with ransomware infected software or content. As these authors mentioned earlier in this text, detecting and avoiding imminent ransomware attacks is still the best defense against a ransomware attack and this methodology achieves the said purpose.

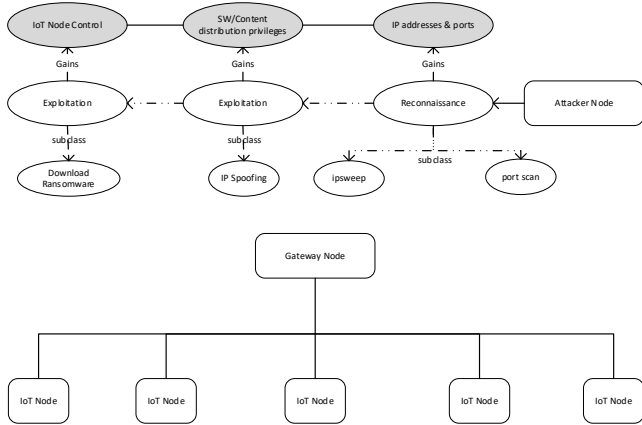


Fig. 2. Building Attack Filters of Ransomware Penetration within IoT Network.

4) *Classification algorithm:* The feature vectors are fed to classifier algorithms such as SVM to find the attack and give an alert as output. SVM modeling algorithm has to find the optimal hyperplane to classify the data. Optimal hyperplane maximizes the margin of training data. The training data set is a set of n elements (x_i, y_i) where x_i is a p dimensional vector, the definition (2) can be given as smallest $\|w\|$ will be giving biggest margin,

$$\text{Minimize in } (w, b) \tag{2}$$

$$\|w\| \text{ subject to } y_i (w \cdot x_i + b) \geq 1$$

(For any $i = 1, \dots, n$)

IV. RESULTS AND DISCUSSIONS

Fig. 3 shows a simplified but typical topological view of a typical industrial IoT network with attack paths/vectors and subsequent events. It involves master node and several endpoint nodes associated with various data acquisition units. Such an entire deployment is usually managed by a dedicated server. As mentioned in the previous section, deployment paths to build predictive models for determining the probability of ransomware attacks have been focused. Subsequent events like port scanning, encryptions, lockouts of data acquisition units are out of scope for our discussion. Hence our testbed is one such network where the master node is leveraged to deploy attacks on endpoints. The master node would typically connect on TCP/IP or BT interface with its endpoints.

Dataset is collected on the testbed in JSON format. An ontology tool is used to get context-aware data to develop attack filters. We used classifier model SVM and training data with context-aware data and without context-aware data is fed into the model. It is tested with the test dataset. In this method,

we overcome the heterogeneity of IoT devices. Context-aware dataset saves time to the tune of 60% compared to the non-aware dataset. Table I presents the time taken by the original dataset and context-aware dataset.

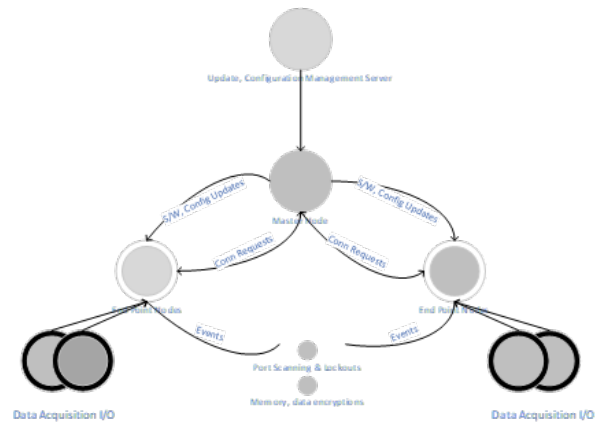


Fig. 3. Topological view of Typical Industrial IoT Network with Attack Paths/Vectors and Subsequent Events.

TABLE I. TIME TAKEN BY ORIGINAL DATASET AND CONTEXT-AWARE DATA SET

Experiment	Time taken by original dataset and context-aware data set.	
	Original dataset	Context-aware dataset
1	3 ms	0.9 ms
2	3 ms	0.8 ms
3	2.8 ms	0.8 ms
4	3.1 ms	1 ms

V. CONCLUSION

A methodology to build prediction models for ransomware attacks on industrial IoTs is developed by focusing on their specific behavior common to most of such devices to overcome challenges posed by their inherent heterogeneity. In this paper, context awareness is used for identifying the most relevant attack paths, vectors, and resultant events to build more effective prediction capabilities.

ACKNOWLEDGMENT

We would like to thank the GITAM University for providing us with the necessary infrastructure for doing this research.

REFERENCES

- [1] A. Giusto, et al., "The Internet of Things," Springer, ISBN: 978-1-4419-1673-0, 2010.
- [2] L. Atzori, et al., "The Internet of Things: A Survey," Computer Networks, Vol. 54, Issue 15, 2787-2805, 2010.
- [3] T. Aliya and L. Wadha, "Security Framework for IoT Devices against Cyber-Attacks," Computer Science & Information Technology (CS & IT), 249-266, 2019.
- [4] Y. Ibrar, et al., "The rise of ransomware and emerging security challenges in the Internet of Things", Computer Networks. Volume 129, Part 2, 444-458, 2017.
- [5] M. U. Kiru and A. B. Jantan, "The Age of Ransomware: Understanding Ransomware and its countermeasures," Artificial Intelligence and

- Security Challenges in emerging networks, R. Abassi, Ed. Pennsylvania: IGI Global, pp. 1–37, 2019.
- [6] A. Wani and S. Revathi, “Ransomware protection in IoT using software defined networking,” *International Journal of Electrical & Computer Engineering*, Vol. 10 Issue 3, 3166-3175, 2020.
- [7] A. Aleroud and K. George, “Contextual information fusion for intrusion detection: a survey and taxonomy,” *Knowledge and Information Systems*, Vol. 52, 563–619, 2017.
- [8] Davidson, et al., “Security Gets Smart with AI,” SANS Institute, 2019.
- [9] H. Martin, et al., “Survey of Attack Projection, Prediction and Forecasting in Cyber Security,” *IEEE Communications Surveys & Tutorials*, 640 – 660, 2018.
- [10] V. Petri and L. Martti, “Artificial intelligence in the cyber security environment,” *The 14th International Conference on Cyber Warfare and Security*, Stellenbosch, South Africa, 2019.
- [11] H. Mahmudul, et al., “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches,” *Internet of Things*, Vol. 7, 2019.
- [12] T. Yuki and F. Satoshi, “Detecting Ransomware using Support Vector Machines,” *Proceedings of the 47th International Conference on Parallel Processing Companion*, Article No.1, pp 1–6, Eugene OR USA, 2018.
- [13] J. Song, et al., “A generalized feature extraction scheme to detect 0-Day attacks via IDS alerts,” *Proceedings of the 2008 international symposium on applications and the internet*, pp 55–61, Turku, Finland, 2008.
- [14] R. Syed, et al., “Ransomware and Internet of Things: A New Security Nightmare”, *9th International Conference on Cloud Computing, Data Science & Engineering*, Noida, India, 2019.
- [15] H. Gavin, et al., “Ransomware deployment methods and analysis: views from a predictive model and human responses”, *Crime Science*, Vol. 8, 2019.
- [16] C. Hong-Yi, et al., “Implementation of ransomware prediction system based on weighted-KNN and real-time isolation architecture on SDN Networks,” *IEEE International Conference on Consumer Electronics*, Taiwan, 2019.
- [17] A. Muna and S. Elena, “Industrial Internet of Things Based Ransomware Detection using Stacked Variational Neural Network,” *Proceedings of the 2019 conference on big data and Internet of Things*, Melbourn VIC Australia, 2019.
- [18] S. Alireza, et al., “A Context-Aware Malware Detection Based on Low-Level Hardware Indicators as a Last Line of Defense,” *SECURWARE 2017: The Eleventh International Conference on Emerging Security Information, Systems and Technologies*, pp 10-19, Rome, Italy, 2017.
- [19] A. Ahmed and K. George, “A Contextual Anomaly Detection Approach to Discover Zero-Day Attacks,” *International Conference on Cyber Security*, Washington, DC, USA, 2012.
- [20] C.E. Shannon, *The Mathematical Theory of Communication*. Univ. Illinois Press, 1971.
- [21] A. Ahmed and K. George, “A System for Cyber Attack Detection Using Contextual Semantics,” *7th International Conference on Knowledge Management in Organizations: Service and Cloud Computing*, Vol. 172, pp 431-442, Baltimore, MD 21250, USA, 2013.
- [22] M. Gergely, and J. Abonyi, “A Review of Semantic Sensor Technologies in Internet of Things Architectures,” *Hindawi Complexity*, Vol.2019,pp1-21,2019.

Contribution to the Improvement of Cryptographic Protection Methods for Medical Images in DICOM Format through a Combination of Encryption Method

Maka Maka Ebenezer¹, Pauné Félix², Malong Yannick³, Simo Ntso Pascal Junior⁴, Nnemé Nnemé Léandre⁵
Department of Computer Engineering and Telecommunications, ENSPD, University of Douala^{1,3,4}
Department of Computer Engineering, ENSET, University of Douala^{2,5}

Abstract—This paper proposes a method for storing and securing medical images in DICOM format. Other methods offered affect the quality of the image. The solution proposed here is based on the AES256 algorithm in Galois/Counter Mode (GCM) which already integrates authentication and signature processes to ensure the integrity of the images manipulated. This solution is implemented by using the Python programming language under the DJANGO framework, libraries such as NUMPHY, PYDICOM, MYSQLCLIENT, and PYCRYPTODOME. The results obtained after experimental tests give us a good average encryption and decryption time. The difference in the mean value of time between encryption and decryption is quite small in view of the tests carried out. We obtain saving on storage space owing to the fact that the proposed solution directly stores the encrypted image. The manipulated image is not altered.

Keywords—Medical images; DICOM; advanced encryption standard (AES); GCM; authentication

I. INTRODUCTION

With the digital evolution, the consumption of intangible goods has significantly increased, resulting in the circulation of large amounts of data on computer networks, in particular on the internet. Distance communication between individuals is growing rapidly and this does not spare the professional field where documents and audio-visual flows are shared. This contributes to the development of services such as teleworking and telemedicine. In the health sector, according to [1], he indicates for example that multimedia Information and Communication Technology (ICT) are likely to provide doctors with decisive help in the search for a better quality of care. Then several other aspects of cybercrime can mar this help.

Questions relating to the protection of the data that is exchanged across the world are increasingly felt despite the security methods put in place. Whether it is Short Message Service (SMS), instant messaging (chat) or electronic messaging (email), this data must remain "Confidential", which means that only authorized people can consult it; "Integral" because they must not undergo any modification by a third person other than the one having the authorization to do so and they must above all remain available. The particular case of the exchange of digital medical images which contain a great deal of information on patients and must imperatively be protected in order to guarantee medical secrecy regulated by Law. To

provide solutions to this issue of digital medical images privacy, several protection methods have been developed to guarantee their Availability, Integrity and Confidentiality (AIC). Among these protection methods, cryptographic ones seem to be the most suitable.

In this paper, we will describe some methods of protecting medical images, and methods of storing medical images. We will propose a cryptography-based solution using a symmetric encryption algorithm combined with an authenticated encryption algorithm designed to provide both data integrity and authenticity, as well as confidentiality (Galois / Counter Mode). The proposed solution will thus improve the storage capacity of medical image backup systems. This will ensure the security of the medical image in Picture Archiving and Communication System (PACS) while ensuring the AIC criterion. We structure this paper into six sections. Section I introduces the methods of cryptographic protection. Section II gives the literature review on the protection of medical images. Section III describes the transmission of medical images. Section IV deals with securing medical image. Section V gives the methodology and results. Section VI concludes this work.

II. PROTECTION OF MEDICAL IMAGES

The protection of medical images is one of the top priority issues about digital image security. Several researchers have focused on work aimed at improving methods of securing the medical image. From the use of watermarking methods for shared medical images [2], in order to ensure the integrity and confidentiality of the data [3], through selective encryption [4] and the use of cryptography-based protection methods on the displacement of Red Green Blue (RGB) pixels [5], without forgetting the algorithms based on the encryption of flows with a function of nonlinear filtering [6].

The medical image has undergone very remarkable changes in recent years thanks to the development of physics. The medical image, by its nature, is supposed to carry a set of information on the patient. Many methods allowing the acquisition of medical images have emerged; for standard IT management of medical imaging data and to ensure interoperability between these different modalities, the DICOM standard has been adopted. DICOM stands for Digital Imaging and Communication in Medicine. The objective of the DICOM standard is to facilitate the transfer of medical images between machines from different manufacturers. It defines a file format

for digital files created during medical imaging examinations as well as a data transmission protocol (based on TCP / IP) [7]. Several techniques are used to ensure the protection of medical images.

A. Cryptographic Protection

Cryptography is the first device to guarantee the security of electronic documents [8]. It allows sensitive information to be stored or transmitted over insecure networks (such as the Internet) so that it cannot be read by anyone other than the intended person. Data that can be read and understood without special measures is called clear text, and the process of hiding clear text so as to hide its substance is called encryption. The operation to recover the clear data from the encrypted data is called decryption. Encryption is usually done using an encryption key, while decryption also requires a decryption key. There are two types of keys, namely symmetric keys, that is to say keys used at the same time for encryption and decryption. This is referred to as symmetric encryption or secret key encryption; and asymmetric keys, which mean that the keys used for encryption and decryption are different, this is referred to as asymmetric encryption or public key encryption.

1) Symmetric key encryption or secret key cryptography:

In this type of system, the same key is shared between the sender and the receiver to encrypt and decrypt information. The problem with this method resides in the secure distribution of the key to the recipient of the encrypted message. Several secret key encryption algorithms have emerged, these include the algorithms of continuous or stream cipher, which act on the clear text and on one bit at a time; block cipher algorithms, which operate on plain text in groups of bits called blocks. And of all these algorithms, according to [8] the most widely used symmetric encryption algorithm is AES. In order to secure the transfer of medical images, William Puech and Develay Morice jointly used the AES algorithm in stream mode and JPEG compression [4].

2) Asymmetric encryption or public key cryptography:

Whitfield Diffie and Martin Hellman invented the concept of public key cryptography in 1976, with the aim of solving the key distribution problem posed by secret key cryptography. Numerous algorithms have emerged for this purpose, all based on sophisticated mathematical problems that are generally difficult to solve. In these algorithms, the encryption and decryption keys are distinct and cannot be deduced from each other. We can therefore make one of the two public while the other remains private. If the public key is used for encryption, anyone can encrypt a message which only the owner of the private key can decrypt. Some algorithms allow the private key to be used for encrpii.

A hash function is a function that will calculate a unique fingerprint (or signature) from the data provided. A cryptographic hash function has some particular characteristics, unidirectional meaning being the most important of them. As a matter of fact, it is a function whose reverse is impossible to

calculate, even by using a great computing power for a long period of time. The most famous according to [4] is Message Digest 5 (MD5) which is still widely used although in terms of security, it is recommended to upgrade to more robust versions because collision suites have been found; this function returns a 128-bit hash. The Secure Hash Algorithm 1 (SHA1) was the replacement function of MD5 because it produced 160-bit hashes with no possibility of finding collisions until 2004-2005, when attacks proved the possibility of generating collisions. Since that date it is no longer recommended to use the SHA1 function. But it is still widely used. We also have SHA256 and SHA512 which are two of the major standards in use today as there have been no attacks so far to detect security holes on these hash functions. They produce signatures of 256 bits and 512 bits, respectively.

B. Data Hiding

Data hiding refers to the insertion into a digital medium of a given quantity of secret binary information imperceptibly and more or less robust, depending on the intended application. The term "concealment" does not mean here that the information is visible but encoded, in this case it would be cryptography. Rather, it means that the presence of the information to be protected (called a useful message) is not perceptible because it is buried in other information (called a cover message). In the case of the protection of digital information, the useful message makes it possible to identify the owner of the cover message or its origin or to guarantee its integrity. Data concealment encompasses two techniques that are very similar to each other, but which do not have the same objectives or the same constraints. Depending on the context, a distinction is made between steganography where it must be impossible to distinguish whether the cover message contains a useful message or not. The most important constraint is then imperceptibility; and digital marking where the useful message is linked to the identity of the beneficiary of the cover document, and must therefore remain present even if the latter undergoes modifications. In this case the main constraint is then robustness [9].

1) Storage of medical images: The reception facilities for digital images are much more accessible in practical terms than those for analogue images [10]. Thanks to the digital storage of medical images, it is possible to comment, view and process them locally or remotely. Image archiving and communication systems PACS are set up for the management of medical images and their communication in the appropriate infrastructures. These systems include an archiving station for long-term storage of image data, and an examination station for displaying images based on received image data [11]. Depending on the needs of the institutions, the storage devices can be local as in the previous case or very often remote thanks to cloud-type infrastructures. In all cases, the short and long term security of the DICOM files for the studies provided from the imaging modalities must be guaranteed. A "study" consists of one or more series of images captured using an imaging modality.

III. TRANSMISSION OF MEDICAL IMAGES

The medical image flow circulates between the source which constitutes the modality (scanner, OTP, CT, etc.) which allows the acquisition of the medical image or any other DICOM image source, then the backup server and archiving (PACS or other system) of imaging files and finally the radiologist's reading workstation.

The transmission of medical images must be done quickly, securely and reliably. This is done by ensuring that all data passing through the network is encrypted. In practice, a VPN / IPSec tunnel is built between the source and the recipient for the routing of data; thus, all transmissions are secure. It is also possible to transmit securely through the DICOM / TLS protocol provided by the DICOM standard, which has been used more and more in recent years; or on the web through SSL / TLS protocol.

As most modern modalities increasingly produce high quality images, this can impact the speed of data transmission which would be a problem for emergency requests. Compression methods are very often used to overcome this problem. DICOM supports lossy and lossless compression mechanisms, such as JPEG2000, RLE, and even JPEG-LS. Other techniques are used to overcome this problem such as parallel data transfer. However, the constant improvement of telecommunications networks and data networks makes it possible to solve this problem without any particular technique.

IV. SECURING MEDICAL IMAGES

The deployment of PACS and electronic medical records requires to significantly increase the security of hospital and radiological information systems in order to ensure the protection of patient's data. The security rules that govern this protection are based on three principles: confidentiality, reliability and availability [12].

Securing a DICOM file consists of four operations, namely: securing access to the file by defining access rights, securing the transmission of the file, digitally signing the file and securing the storage [13]. Assuming that a secure transmission has been established, we now need to ensure that stored medical images are protected and that processes and procedures are in place to ensure data security and availability only to authorized users. The proposed work fully contributes to the management of access control, digital signature and secure storage of files. It combines the Hash Message Authentication Code (HMAC), RSA and AES methods to ensure the confidentiality and integrity of images.

A. AES Algorithm

This algorithm was officially approved as a standard on December 6, 2001 [14]. It is a block cipher algorithm for encrypting a clear text consisting of 128 bits of data using a secret key consisting of 128, 192 or 256 bits. Its operating principle consists in taking as input a block of 128 bits, the key being 128, 192 or 256 bits. The 128 input bits are "mixed" according to a previously defined table. These bytes are then

placed in a 4x4 square matrix. Line items are rotated to the right. The increment for the rotation varies depending on the row number. A transformation is then applied to the matrix by an XOR with a key matrix. Finally, an XOR between the matrix and another matrix makes it possible to obtain an intermediate matrix. These different operations are repeated several times and define "one turn". For a key of 128, 192 or 256, AES requires 10, 12 and 14 turns, respectively, depending on the size of the key.

B. Message Authentication Code

Better known by the acronym MAC, Message Authentication Code, these are cryptographic functions intended to verify the integrity of data and to authenticate its origin. These MACs work in a similar way as hash functions. They calculate from a message of arbitrary length a summary of fixed length (this summary is called a hash). But, unlike hash functions, this summary also depends on a secret key K .

A message authentication code is a function $h(M, K)$ where M is the message and K the key, which returns fixed-length text. The calculation of $h(M, K)$ must be done very quickly; and if we know examples of code calculated with the same key, say $h(M_1, K), \dots, h(M_n, K)$ and if we have a new message M , but not the key K , we cannot calculate $h(M, K)$.

When communicating between individuals, MACs are used in the following way: a sender and recipient begin by agreeing on a secret key, through a secure channel. The sender, when he wants to send a message, calculates his MAC using the secret key, and jointly sends the message and his MAC code. On arrival, the recipient also calculates the MAC using its own secret, and compares it with the version sent. If the two coincide, he is sure of both who sent him the message and also that this message has not been modified. Otherwise, the integrity of the message is compromised.

C. RSA Algorithm

It is one of the most popular asymmetric encryption algorithms. Its principle is based on the problem of factoring large numbers. It is extremely difficult to set up a fast algorithm capable of finding two prime numbers whose product is a known number. This is even more difficult when the numbers used are very large. Suppose that an Entity A wishes to send a message to an Entity B. It makes a communication request to Entity B. The key creation step is the responsibility of Entity B. It does not intervene at each encryption because the keys can be reused. The first difficulty which encryption does not solve, is that Entity A is quite certain that the public key it holds is that of Entity B. The renewal of the keys only occurs if the private key is compromised, or as a precaution after a certain time (which can be counted in years). The principle of key creation is as follows:

1. Choose p and q , two distinct prime numbers;
2. calculate their product $n = p \times q$, called the encryption module;
3. calculate $\phi(n) = (p-1)(q-1)$ (it is the value of the indicatrix of Euler in n);
4. choose a natural number e prime with $\phi(n)$ and strictly less than $\phi(n)$, called the encryption exponent;
5. calculate the natural number d , inverse of $e \equiv \phi(n)$, and strictly less than $\phi(n)$, called the decryption exponent; d can be calculated efficiently by the extended Euclidean algorithm.

As e is prime with $\phi(n)$, according to the Bachet-Bézout theorem there are two integers d and k such that $ed = 1 + k\phi(n)$, that it to say that $ed \equiv 1 [\phi(n)]$ and e is actually invertible modulo $\phi(n)$.

The pair (n, e) or (e, n) [15] is the public key of the encryption, while its private key is [16] the number d , knowing that the decryption operation requires only the private key d and the integer n , known to the public key (the private key is sometimes also defined as the pair (d, n) [15] or the triplet (p, q, d)).

If M is a natural number strictly less than n representing a message, then the encrypted message will be represented by $M^e \equiv C[n]$ and the natural number C being chosen strictly less than n .

To decipher C , we use d , the inverse of $e \text{ mod } (p-1)(q-1)$, and we find the plain message M by $M = C^d [n]$.

V. METHODOLOGY AND RESULTS

The solution implemented in this paper is based on a tool for sharing and archiving medical images. To carry out this work we used tools such as PYTHON which is an interpreter, multi-paradigm and multiplatform programming language. The python DJANGO framework is for the development of web applications. The NUMPY library of the Python programming language, intended to handle multidimensional matrices or arrays as well as mathematical functions operating on these arrays. The PYDICOM library of the Python language allowing the manipulation of DICOM files. The MYSQLCLIENT library of the Python language is for the connection to the MySQL database manager. The PYCRYPTODOME library of the stand-alone Python language

of low-level cryptographic primitives and the Radian DICOM viewer-2 software is for viewing medical images.

DICOM files are made up of image details and patient details. These files are organized in tags referring to specific information. Fig. 1 shows an extract from the display of the tags of a DICOM file via the PYDICOM library from the command `pydicom.dcmread('file name.dcm')`.

```
Dataset.file_meta -----
(0002, 0000) File Meta Information Group Leng
(0002, 0001) File Meta Information Version
(0002, 0002) Media Storage SOP Class UID
(0002, 0003) Media Storage SOP Instance UID
(0002, 0010) Transfer Syntax UID
(0002, 0012) Implementation Class UID
(0002, 0013) Implementation Version Name
(0002, 0016) Source Application Entity Title
-----
(0008, 0005) Specific Character Set
(0008, 0008) Image Type
(0008, 0012) Instance Creation Date
(0008, 0013) Instance Creation Time
(0008, 0016) SOP Class UID
(0008, 0018) SOP Instance UID
(0008, 0020) Study Date
(0008, 0022) Acquisition Date
(0008, 0023) Content Date
(0008, 0030) Study Time
(0008, 0032) Acquisition Time
(0008, 0033) Content Time
(0008, 0060) Modality
(0008, 1030) Study Description
```

Fig. 1. Visualization of some DICOM Tags.

Fig. 2 and Fig. 3 respectively show the process of encryption and decryption of the proposed process.

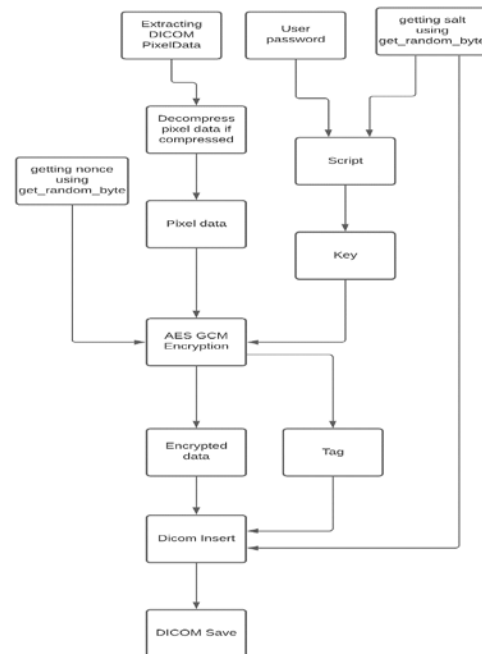


Fig. 2. Diagram of an Image Encrypting Process in DICOM Format.

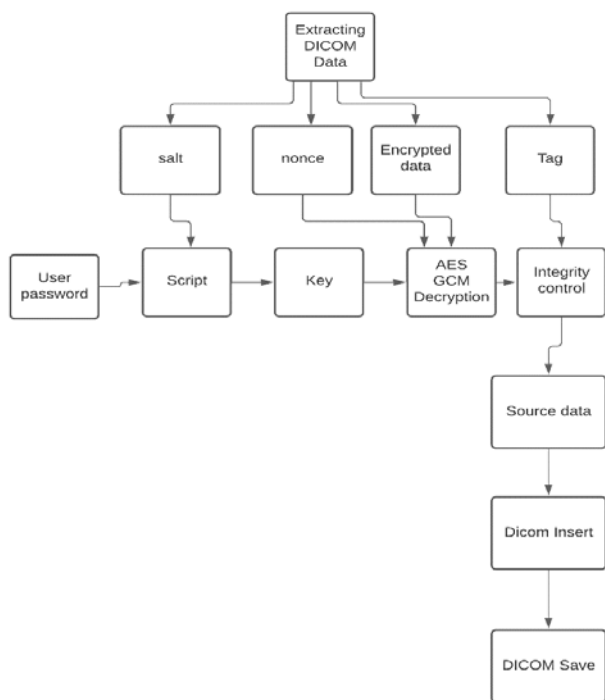


Fig. 3. Diagram of the Process of Deciphering a Cipher.

A. Application of the Encryption Process for Medical Images with the AES Method in GCM Mode

The encryption method adopted is to encrypt the entire contents of the file. A 32-byte encryption key is obtained by calculating the hash of a password from the script function. Then this key is used to initialize a cipher block from a 16-byte initialization vector. The chosen encryption mode is GCM.

Fig. 4 shows the extraction of DICOM tags before encryption. We can very well observe the corresponding values for each element concerning the patient.

```
Dataset.file.meta -----
(0002, 0000) File Meta Information Group Length  UL: 202
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID       UI: Secondary Capture
(0002, 0003) Media Storage SOP Instance UID    UI: 1.2.276.0.7230014
(0002, 0010) Transfer Syntax UID               UI: JPEG Baseline (P
(0002, 0012) Implementation Class UID         UI: 1.2.276.0.7230014
(0002, 0013) Implementation Version Name      SH: 'OFFIS_DCMTK_360
-----
(0008, 0005) Specific Character Set            CS: 'ISO_IR 100'
(0008, 0016) SOP Class UID                     UI: Secondary Capture
(0008, 0018) SOP Instance UID                  UI: 1.2.276.0.7230014
(0008, 0020) Study Date                       DA: '19010101'
(0008, 0030) Study Time                       TM: '000000.00'
(0008, 0050) Accession Number                 SH: ''
(0008, 0060) Modality                         CS: 'CR'
(0008, 0064) Conversion Type                  CS: 'WSD'
(0008, 0090) Referring Physician's Name       PN: ''
(0008, 103e) Series Description                LO: 'view: PA'
(0010, 0010) Patient's Name                   PN: 'ff8563d5-cc01-4d
(0010, 0020) Patient ID                       LO: 'ff8563d5-cc01-4d
(0010, 0030) Patient's Birth Date             DA: ''
(0010, 0040) Patient's Sex                   CS: 'F'
(0010, 1010) Patient's Age                    AS: '37'
(0018, 0015) Body Part Examined              CS: 'CHEST'
(0018, 5101) View Position                   CS: 'PA'
(0020, 0000) Study Instance UID               UI: 1.2.276.0.7230014
```

Fig. 4. Extract from DICOM tags before Encryption.

It should be remembered that bits as we have specified above, represent the content of the DICOM file. Fig. 5, Fig. 6 and Fig. 7 show on the first line (on the left the real image and on the right the corresponding histogram); on the second line, we have on the left the numbered image and on the right its corresponding histogram. This line clearly shows that the encryption process is indeed effective and its corresponding histogram sufficiently reflects the difference with the histogram of the real image. The third line reflects the result of the decryption process and we can observe the stability of these images compared to those of the first line.

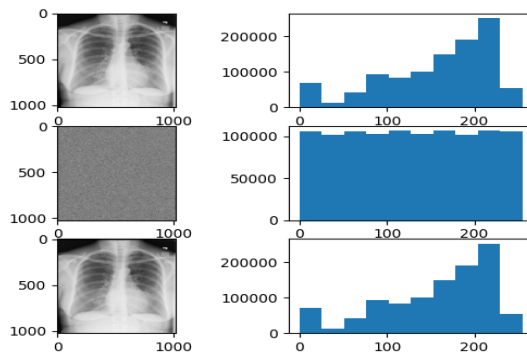


Fig. 5. Result of the Encryption and Decryption Process of a Radiology.

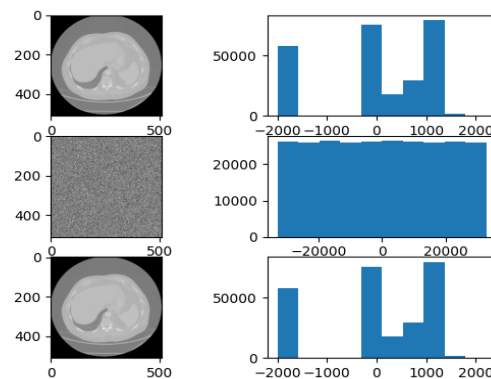


Fig. 6. Result of the Process of Encryption and Decryption of a Tomography.

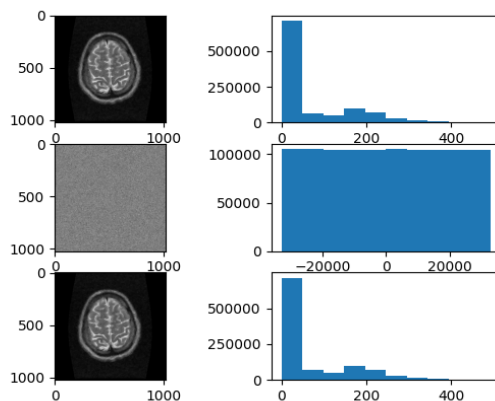


Fig. 7. Result of the Encryption and Decryption Process of an IRM.

B. Signature of the DICOM File

The owner of the image in the system has an RSA key pair. Image data such as: patient name, image pixels, patient ID; are combined and then compressed using the SHA256 function. The hash thus obtained is encrypted using the owner's private key and inserted into a tag in the DICOM file: the tag (0400,0120) or Signature.

To check the integrity of an image, the same data is extracted and the hash is recalculated. The signature is decrypted using the owner's public key accessible to everyone. The hashes thus obtained are compared. Since the hash functions are strict, the slightest change in the image will result in a drastic change in the hash. We can therefore say that the integrity of the image is preserved.

C. Evaluation of the Response Time of the Encryption and Decryption Process

The tests carried out on a radiology-type modality for a DICOM file of 4,739,280 bytes, i.e. 4.5MB, made it possible to obtain the Table I, which lists a set of 10 tests on the same medical image, thus making it possible to obtain each time the encryption and decryption time of said image.

Table I shows a variation of the encryption and decryption time. Note that the average encryption time for the tests performed is approximately 21.82 ms with a minimum time of 21.38 ms and a maximum time of 23.17 ms for a sample of 10 tests. Regarding the decryption, the average time is 22.12 ms for a minimum time of 21.09 ms (that is to say near to the minimum encryption time) and for a maximum time of 23.99 ms (near to the maximum encryption time). From the above results, we can attest to the speed of the encryption and decryption process of the AES algorithm in GCM mode.

TABLE I. EVALUATION OF THE ENCRYPTION AND DECRYPTION TIME OF THE PROCESS IMPLEMENTED IN A RADIOLOGY MODALITY

	Encryption Time (ms)	Decryption Time (ms)
T1	23,17	23,99
T2	21,53	21,33
T3	21,38	21,21
T4	21,53	21,48
T5	21,53	21,94
T6	21,58	21,91
T7	22,04	22,19
T8	21,83	21,09
T9	21,84	22,63
T10	21,81	22,50
MAX	23,17	23,99
MIN	21,38	21,09
AVG	21,82	22,12

VI. CONCLUSION

In this paper, we have proposed an efficient solution which is both, reliable and fast using a mechanism to reinforce the security of the contents of DICOM files in the PACS. To achieve this, we combined the AES symmetric encryption algorithm with the GCM authenticated encryption algorithm. As part of this work, the DICOM file is fully encrypted and then stored. The proposed solution requires less storage space in PACS because the content of the image is directly encrypted, the properties of the file remain the same. According to the experimental results, the image quality is not affected. The time required for encryption is on average 21.82 ms and that for decryption is 22.12 ms. Given these results, we can say that the minimum encryption time is equal to the average encryption time to the unit; the same is true for the decryption time. It should also be noted that the average difference in absolute value of the encryption and decryption time is of the order of 0.3 ms.

REFERENCES

- [1] Alain Venot, "Security, legal and ethical aspects of computerized health data," in Medical Informatics, e-Health, Paris, 2013.
- [2] M. Karasad, "Tattooing Shared Medical Images," p. 164.
- [3] M. A. Hajjaji, H. Ridha, M. Abdellatif, and B. El-Bey, "Tattooing Medical Images for Data Integrity and Privacy," Tunisia, nov. 2010, Accessed: Oct 24, 2020. [Online]. Available at: <https://hal.archives-ouvertes.fr/hal-00822661>.
- [4] W. Puech, J. Rodrigues, and J.-E. Develay-Morice, "Secure Transfer of Medical Images by Joint Coding: Selective Encryption by AES in Stream Mode and JPEG Compression," nov. 2006.
- [5] Q.-A. Kester, "Image encryption based on the RGB PIXEL transposition and shuffling," Int. J. Comput. Netw. Inf. Secur., vol. 5, p. 43-50, june 2013, doi: 10.5815/ijenis.2013.07.05.
- [6] Belmeguenai Aïssa, Derouiche Nadir and Mansouri Khaled, "Security analysis of image cryptosystem using stream cipher algorithm with nonlinear filtering function," International Journal of Advanced Computer Science and Applications(IJACSA),3(9),2012.
- [7] Marie-Hélène Coste and Véronique Simon, "Journey to the Heart of Medical Imaging Networks," Press kit, French Society of Radiology.
- [8] E. Coumet, "Cryptography and numeration," Ann. Hist. Sci. Soc., vol. 30, no 5, p. 1007-1027, 1975.
- [9] Chikhi Samia Boucherkha, "Contribution to the Flexible Authentication of Digital Images Using Image Marking Techniques: Application to Medical Images," oct. 2008.
- [10] J.S. DELMOTTE and G. GAY, "Modern medical imaging applied to internal medicine, technical and practical aspects," Lille, Nancy (France).
- [11] R. E. C. Jr, M. G. Gaeta, D. M. Kaufman, et J. G. Henrici, "Picture archiving and communication system," US6574629B1, june 03, 2003.
- [12] Romain Héroult, "Image tattooing and cryptography: To ensure the confidentiality, integrity and authentication of medical images and to insert confidential data," DEA practical internship, August 20, 2004.
- [13] P. Subhasri et D. A. Padmapriya, "Enhancing the security of dicom content using modified vigenere cipher," vol. 10, p. 7, 2015.
- [14] FIPS 197, "Advanced Encryption Standard (AES)," nov. 2001.
- [15] « Rivest, Shamir et Adleman 1978, p. 122. ».
- [16] « Menezes, van Oorschot et Vanstone 1996, chap. 8, p. 286 ; Schneier 1996, Applied Cryptography, p. 467. ».

Birds Identification System using Deep Learning

Suleyman A. Al-Showarah¹, Sohyb T. Al-qbailat²

Faculty of Information Technology
Mutah University, Karak
Jordan

Abstract—Identifying birds is one of challenging role for bird watchers due to the similarity of the birds' forms/image background and the lack of experience for watchers. So, it needs a computer system based images to help birdwatchers in order to identify birds. This study aims at investigating the use of deep learning for birds' identification using convolutional neural network for extracting features from images. The investigation was performed on database contained 4340 images that collected by the paper author from Jordan. The Principal Component Analysis (was applied on layer 6 and 7, as well as on the statistical operations of merging the two layers like: average, minimum, maximum and combine of both layers. The datasets were investigated by the following classifiers: Artificial neural networks, K-Nearest Neighbor, Random Forest, Naïve Bayes and Decision Tree. Whereas, the metrics used in each classifier are: accuracy, precision, recall, and F-Measure. The results of investigation include and not limited to the following, the PCA used on the deep features does not only reduce the dimensionality, and therefore, the training/testing time is reduced significantly, but also allows for increasing the identification accuracy, particularly when using the Artificial Neural Networks classifier. Based on the results of classifiers; Artificial neural networks showed high classification accuracy (0.9908), precision (0.718), recall (0.71) and F-Measure (0.708) compared to other classifiers.

Keywords—Birds identification; deep learning convolutional neural networks (CNN); VGG-19; principal component analysis (PCA)

I. INTRODUCTION

Many people are interested in observing and studying wildlife, especially in birdwatching. The role of birdwatching is to preserve the nature by observing bird's behavior and migration pattern. The challenge for bird watchers in identifying birds based images remains difficult due to the similarity of the birds' forms/ image background and the lack of experience in this field for watchers [1].

As mentioned in [17] that birds Voice or Videos were used in earlier technique to predict it species, but this technique have many challenges to give an accurate result due to other background of birds/animal voices. So, images can be best choice to be used to identify birds' species. To implement this technique, the images for all birds' species need to be trained to generate a model. Then deep learning algorithm will convert uploaded image into gray scale format and apply that image on train model to predict best match species name for the uploaded image.

Also, during the previous years, artificial intelligence is used in the field of bird watching based images using different

algorithms and methods [1][3][4][7][14], but this study differs from others in using the following operations: combine between the fc6/fc7, max between fc6/fc7, min between fc6/fc7, and the average for fc6/fc7 based on VGG-19. Hence, the field of birdwatching needs more investigations to develop systems with new technique that help to identify birds.

As the database of images were collected from Jordan, and the statistics number of birds in Jordan as stated in [13] are 434 species belonging to 66 families.

This study aims at investigating the use of deep learning for birds' identification using VGG-19 for extracting features from images. In order to achieve this aim, the investigation for the performance of different classifiers were performed on the following classifiers: (KNN, Decision Tree, Random Forest, and ANN) on the collected reliable database of birds images that available in Jordan.

VGG-19 considered as one of the most important models of Convolutional Neural Networks (CNN). Therefore, CNN is considered as the strongest technique for deep learning used in image identification [9].

The main reason of using VGG-19 is to provide high precision by finding features with distinctive details in the image like the difference in lighting conditions and other objects surrounding the birds [3]. Moreover, PCA could be employed as dimensionality reduction tools with these features that would help to reduce number of features that will make the training time less.

The motivation to conduct this study represented by: 1) The shortage in the field of identifying birds based on images. 2) To the best of our knowledge, we have not come across to any study conducted using VGG-19 for identifying birds. 3) There is shortage in database available in the world except these two databases that available in [1] [18]. This case is applicable to Jordan, as there is no database of images for birds, and there is no program was developed to identify birds.

Based on the extracted features using VGG-19, the contribution of this study can provide a research fields with a comparison between the results of different aforementioned classifiers.

This study organized into six sections. Section II introduces the overview of previous studies on all related subjects. Section III describes the used database. Section IV discusses the model design and the methodology for the experiment. Then Section V discusses the results of the experimental, and finally, Section VI presents paper conclusion.

II. RELATED WORK

Machine learning (ML) represents a set of techniques that allow systems to discover the required representations to features detection or classification from the raw data. The performance of works in the classification system depends on the quality of the features. As such of this study can be categorized under the field of ML; this is to make a search in this area for the studies that belong to birds' identification.

In the literature review, there are number of studies conducted in field of identifying birds. But they were conducted in different algorithms and methods, as follows:

There are number of studies conducted for identifying birds based audio/ video like [4][11][6][10]. While other studies conducted to identify birds based images using AI algorithms [1][3][14], but not in what was conducted in this study. This study used different operations like: MAX, MIN, AVERAGE, and Combine between the layers fc6/fc7 based on VGG-19 algorithm.

In field of birds database-based images and birds identification system, the researchers in [19] conducted study on data collected mostly from North American of 200 bird species, where they called it: (Caltech-UCSD Birds 200 (CUB-200)). They conducted their study based on two simple features: image sizes and color histograms. In the case of image sizes, they represented each image by its width and height in pixels. But in case the color histograms, they used 10 bins per channel, where an applied Principal Component Analysis was applied. Their results showed how the performance of the NN classifier degrades as the number of classes in the dataset is increased, as in [18]. The performance of the image size features are close to chance at 0.6% for the 200 classes, while the color histogram features increase the performance to 1.7%. Another example of studies that conducted in field of database for birds based images and birds' identification system, the researchers in [18] increased the number of images to 11788 images; as it was 6033 in [19]. Where they used RGB color histograms and histograms of vector-quantized SIFT descriptors with a linear SVM. The results obtained of their study for the classification accuracy is 17.3%.

Also, in the field of birds' identification system, the researchers in [14] proposed a new feature to distinguish the types of birds. In their study, they used the ratio of the distance from the eye to the beak root, and the beak width. This feature was integrated in the decision tree, and then in SVM. This proposal was applied to the database that called (CUB-200-2011 dataset) that mentioned in [18]. The results achieved for correct classification rate is 84%.

Another study conducted on birds-identification. Their database was collected in India by the researchers that

available in [1]. In their study, their database consisted of 300-400 different images consists of number of bird species. In their study, the algorithm used to extract image features is AlexNet and then classified by using a SVM classifier. The results of accuracy is 85%.

The researchers in [11] used multiple pre-CNN networks algorithms like: (AlexNet, VGG19 and GoogleNet) on birds dataset that is called (Caltech-UCSD Birds-200-2011). Based on approach of combining between the aforementioned algorithms together, the results showed that this approach improved the accuracy that reached to 81.91%, when applied on Caltech-UCSD Birds-200-2011 dataset compared to other datasets used in the same study.

Another study conducted by [4] in field of database birds based images and birds identification system. Their study aimed to classification the birds during flight from video clips. They approximately collected 952 clips and extracted about 16,1907 frame photos of 13 birds' species. In order to improve the accuracy, the researchers used the two features: appearance and motion features. Then, they compared their proposed method with the classifiers (VGG, MobileNet). The proposed method achieved a 90% correct classification rate when using Random forest classifier.

In field of birds' identification system, the researchers in [3] applied different methods like: 1) softmax regression using manually features on the Caltech-UCSD-Birds-200 dataset [19]. 2) A multi-class SVM was applied on HOG and RGB on features extracted from images. 3) A CNN was applied using transfer learning algorithm to classify birds. The results of comparing the three methods 46% when using CNN.

In the next section, the database content, number of images, source of images, and the challenges to classify images are explained.

III. DATABASE DESIGN

The database of birds images were collected from Jordan, and it consists of 4340 images of 434 bird species. The database images were obtained from scientific sources and were approved by Jordanian Bird Watching Association based on their scientific names [13].

The images have different backgrounds, where some of them were taken in shadow condition, lightening background, and some of them have other objects in the images as background. This has added a huge challenge to the researchers to extract features, and to provide high accuracy.

IV. PROPOSED METHOD

This section presents the procedures that used for the proposed method in identifying birds using VGG-19. Fig. 1 shows the proposed model.

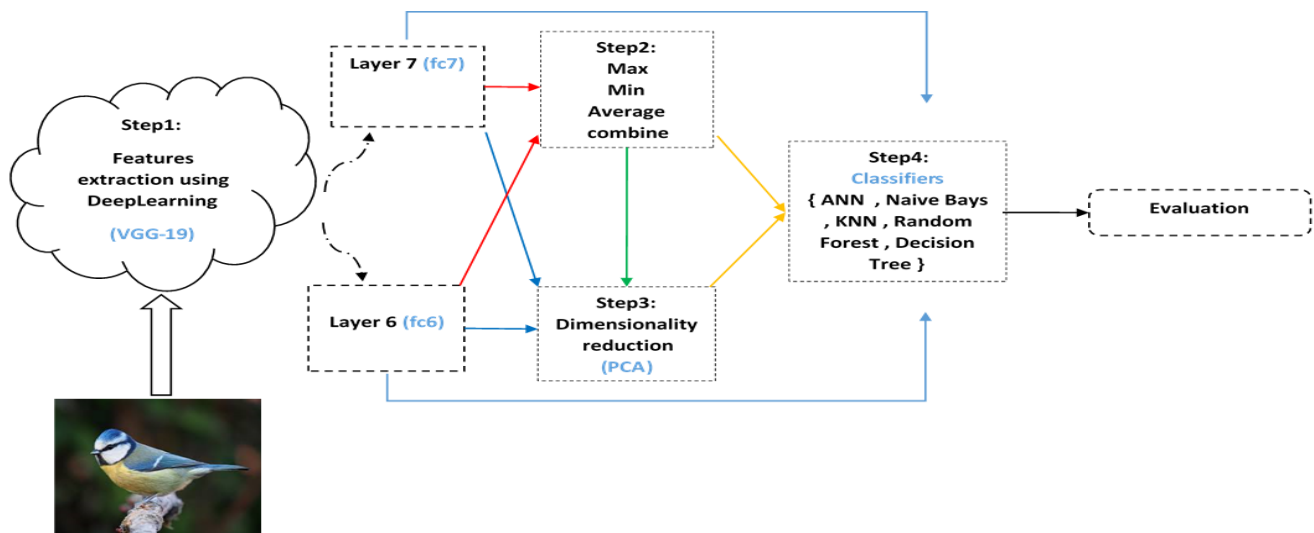


Fig. 1. Proposed Model.

The following steps explain the proposed model of this study, as follows:

Step 1): The feature vectors will be extracted from images automatically using MATLAB for Pretrained VGG-19 to build dataset that includes (feature factors: fc6 and fc7). Each dataset (e.g. fc6) contains 4096 columns (representing feature vectors) and 4340 rows (representing the number of samples (images)).

Step 2): The statistical operations like: (min, max, average, and combined them together) were performed on the original/pure of fc6 and fc7 layers, this is to obtain new dataset to be used in the next stage (step 3) of using classifiers. Explanation on statistical operations, as follows:

- Max: used to find the largest value between the two values in fc6 and fc7 and put value in a new group.
- Min: used to find the less value between the two values in fc6 and fc7 and put value in a new group.
- Average: used to find average the two values in fc6 and fc7 and put value in a new group.
- Combined them together: used to combine the first group (4096) next to the second group (4096). This is to have a new group that contains 8192 features in this study.

Step 3): A PCA will be applied on the original/pure of fc6, fc7, the dataset that obtained from the previous stage (step 2); this is to produce a new datasets.

The data obtained using the pre-trained VGG-19, is very large (4096), therefore, the PCA was implemented to reduce the number of features. In PCA, there were set of percentages used to show the variance of the data in the results, which are: 95%, 97% and 99% variance of the data (the 4096 features).

Step 4): The results were performed based on applying set of classifiers on the datasets that obtained from (step 2 and step 3).

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section presents the performance evaluation results for the study dataset, which include the accuracy, F-measure, recall, precision and training time for each classifier as follows:- 1KNN, 3KNN, 5KNN, ANN, Naive Bayes, Random Forest and Decision Tree.

The results of this study are displayed as follows:

A. Results of both Original/Pure fc6/fc7 Datasets Separately

Table I shows the results of both original of fc6 and fc7 datasets. Naive Bayes has achieved the highest accuracy results for fc6 and fc7 which are (59.002) and (56.106).. While for the time spend to conduct the test and training dataset, Decision Tree has spend large time (1406.69s), but KNNs spend less time (0s) compared to other classifiers. This is because it has no training model; where the test example is compared directly to other examples in the training set, and that why it is slow in testing, particularly when used a large number of examples in the training [8][16]. This results match with the results in [5] [12].

B. Results of the Statistical Operations on fc6 and fc7 Datasets

The section show the results of three dataset by applying statistical operations (average, maximum, minimum) between the fc6 and fc7 layers.

Table II shows results of the statistical operations on fc6/fc7 datasets, where Naive Bayes has achieved the highest accuracy results for AVERAGE, MAX, and MIN, which are (57.30), (60.99), and (57.60) respectively. Despite of the Naive Bayes have scored acceptable accuracy, F-measure, recall, and precision that outperformed all classifiers, but also it was achieved with acceptable training time. This result mismatch with other studies [2] [15].

C. Results of Combine between (Original fc6/ fc7) Dataset

A new dataset was obtained called combine by combining of fc6 (4096) and fc7 (4096), which contained 8192 feature vector, and accordingly will obtained the results:

Table III shows the birds identification results where Naive Bayes has achieved the highest accuracy results in combine (59.4009) of accuracy. The second high result of accuracy is 1KNN that has achieved accuracy of 50.2074. While for the time spend to conduct the test and training dataset, Decision Tree spend large time (12484.01s), but KNNs spend less time (0s) compared with other classifiers.

D. Results of Both Original/pure fc6/fc7 after Applying PCA

Tables IV to V shows the identification results for each classifier after applying PCA (95%,97%,and 99%).

In Table IV, the classifier ANN was not used in the previous Tables I to III. This can be explained as follows: ANN is the best classifier to be used for deep features, if and only if it is provided with a smaller number of deep features, otherwise, i.e. if it is applied on the original/pure deep features, which obtained from the VGG-19 layer 6 or 7 or any merging of them both, the training time would be unacceptably long [2] [15][11].

TABLE I. IDENTIFICATION RESULT OF BOTH ORIGINAL/PURE FC6/FC7 DATASETS

Classifiers	Accuracy	Precision	Recall	F-measure	Training Time (seconds)
fc6 results					
1KNN	47.0507	0.542	0.471	0.479	0
3KNN	41.7512	0.506	0.418	0.421	0
5KNN	44.3318	0.535	0.443	0.451	0
Naïve Bayes	59.0092	0.642	0.59	0.601	9.15
Random forest	14.447	0.227	0.144	0.153	35.93
Decision Tree	12.8802	0.133	0.129	0.127	1438.85
fc7 results					
1KNN	50.8065	0.552	0.508	0.511	0
3KNN	46.1751	0.544	0.462	0.463	0
5KNN	47.4885	0.556	0.475	0.48	0.02
Naïve bayes	56.106	0.609	0.561	0.571	8.23
Random forest	21.2442	0.295	0.212	0.22	60.53
Decision Tree	17.5115	0.185	0.175	0.174	1406.69

TABLE II. IDENTIFICATION RESULTS OF AVERAGE, MAXIMUM AND MINIMUM (FC6⊕FC7)

Classifiers	Accuracy	Precision	Recall	f-measure	Training Time (sec)
Avg results					
1KNN	47.0276	0.536	0.47	0.478	0
3KNN	41.5668	0.497	0.416	0.418	0
5KNN	43.8479	0.523	0.438	0.444	0
Naïve Bayes	57.3041	0.624	0.573	0.584	7.11
Random forest	14.424	0.22	0.144	0.148	44.96
Decision tree	13.3641	0.139	0.134	0.131	1278.55
Max results					
KNN	49.6313	0.577	0.496	0.505	0
3KNN	44.9309	0.555	0.449	0.456	0.11
5KNN	47.5806	0.583	0.476	0.487	0
Naïve Bayes	60.9908	0.67	0.61	0.622	7.08
Random forest	16.8433	0.265	0.168	0.176	31.28
Decision tree	14.9078	0.154	0.149	0.148	1467.56
Min results					
1KNN	44.9309	0.513	0.449	0.456	0
3KNN	39.2627	0.491	0.393	0.396	0.02
5KNN	40.1152	0.494	0.401	0.408	0
Naïve Bayes	57.6037	0.632	0.576	0.586	7.19
Random forest	12.9493	0.204	0.129	0.133	53.8
Decision tree	11.0829	0.118	0.111	0.11	1198.16

TABLE III. IDENTIFICATION RESULTS OF COMBINE BETWEEN (ORIGINAL FC6/ FC7) DATASET

Classifiers	Accuracy	Precision	Recall	f-measure	Training Time (sec)
Combine results					
KNN	50.2074	0.555	0.502	0.506	0.01
3KNN	45.2995	0.539	0.453	0.455	0.74
5KNN	47.6728	0.563	0.477	0.482	0
Naïve Bayes	59.4009	0.64	0.594	0.603	14.68
Random forest	18.1797	0.256	0.182	0.185	47.38
Decision tree	16.129	0.166	0.161	0.159	2484.01

TABLE IV. IDENTIFICATION RESULTS OF ORIGINAL/PURE FC6 AFTER APPLYING PCA (95%,97%,99%)

Classifiers	Accuracy	Precision	Recall	F-measure	Training Time (sec)
fc6 (PCA 95%) results					
ANN	68.8018	0.695	0.688	0.685	23378.32
1KNN	27.6959	0.52	0.277	0.317	0
3KNN	15.2074	0.391	0.152	0.185	0
5KNN	15.2765	0.416	0.153	0.19	0
Naïve Bayes	52.0737	0.631	0.521	0.549	0.41
Random forest	6.7281	0.13	0.067	0.067	20.99
Decision tree	14.5392	0.153	0.145	0.144	107.12
fc6 (PCA 97%) results					
ANN	70	0.658	0.65	0.648	19022.88
KNN	19.2857	0.49	0.193	0.237	0
3KNN	8.1797	0.278	0.082	0.104	0
5KNN	8.1567	0.292	0.082	0.107	0
Naïve Bayes	48.318	0.622	0.483	0.52	1.11
Random forest	3.8018	0.085	0.038	0.038	27.03
Decision tree	14.1014	0.154	0.141	0.141	188.18
fc6 (PCA 99%) results					
ANN	62.3733	0.642	0.624	0.623	48850.24
KNN	8.6636	0.325	0.087	0.113	0
3KNN	1.8433	0.072	0.018	0.022	0
5KNN	1.9355	0.079	0.019	0.023	0
Naïve bayes	37.9032	0.581	0.379	0.428	1.44
Random forest	2.0507	0.04	0.021	0.02	28.25
Decision tree	13.1567	0.143	0.132	0.132	471.77

Applying PCA has influenced on the training time for fc6 that made it less for all classifiers in Table IV-after applying PCA compared to the training times in Tables I to III, before applying PCA, especially for Random Forest and Naïve Bayes. The highest accuracy resultant from applying PCA of (95%, 97% and 99%) is in favor of ANN with (68.8018, 70 and 62.3733%), respectively, which can be attributed to the reduced feature vector.

So, it is worth mentioning that the ANN classifier was not used with other sets except those obtained after applying the PCA, this is because of its unacceptable training time. This results matches with previous studies that stated the training time for ANN spend large compared with other classifiers [2][15].

Table V shows the birds identification results for fc7 where the highest accuracy resultant from applying PCA of (95%,97% and 99%) are in favors of ANN with (65.2995, 65.2995 and 67.9493), respectively.

The second high accuracy resultant from applying PCA of all percentage of (95%, 97% and 99%) is Naïve Bayes, has achieved accuracy of (58.3641, 56.9585 and 56.3825%), respectively.

E. Results of the Statistical Operations on (fc6 and fc7) after Applying PCA

This section presents the identification results of the statistical operations on each of (average, maximum and minimum) between the fc6 and fc7 after applying PCA

(95%,97%,99%), as well as the results of training time for each classifier, as follows:

Table VI shows the birds identification results in (average between (fc6 and fc7)) where the highest accuracy resultant from applying PCA of (95%, 97% and 99%) are in favors of ANN with (69.5622, 69.9078 and 65.5069) respectively. The second-high accuracy resultant from applying PCA of all percentage of (95%, 97% and 99%) is Naïve Bayes that has achieved accuracy of (53.3871, 49.7926 and 39.8157%) respectively. While the time spend to conduct the test and training dataset, ANN spend large time 58379.22s , where that PCA 95 spend less time compared to PCA 97and PCA99.

Table VII shows the birds identification results in (maximum between (fc6 and fc7)) where the highest accuracy resultant from applying PCA of (95%) are in favors of ANN with (66.9816) . It is noted that the results of the ANN is appeared only for PCA (95%), but not for the percentage of (97%, and 99%). This is because the large number of features for each of PCA (97% and 99%) that reached to (1428, and 2117) features, respectively. Therefore there will not be results when using ANN, due to its unacceptable training time (that takes days to provide the results.

While for the time spend to conduct the test and training dataset, ANN spend large time 54151.88s.

Table VIII shows the birds identification results in (minimum between (fc6 and fc7)) where the highest accuracy resultant from applying PCA of (95%) are in favors of ANN with (70.8295). It is noted that the result of the ANN is appeared only for the PCA (95%), but not for the percentage (97%, and 99%). This is because the large number of features for each of PCA (97% and 99%) that reached to (1205 and 1910) features respectively. Also, due to its unacceptable training time (that takes days to provide the results. While Naïve Bayes achieved accuracy resultant from applying PCA of all percentage of (95%, 97% and 99%), they are as follows (48.7327, 44.1014 and 35%), respectively. While for the time spend to conduct the test and training dataset, ANN spend large time 42677.02s.

F. Results of Combining Feature Vector after Applying PCA

This section shows the results of combining between fc6 (4096) and fc7 (4096) that reached 8192, but this number of features have been reduced after applying PCA (95%, 97%, 99%) that become (250, 440 and 1080) features, respectively. The results of combine, as follows:

TABLE V. IDENTIFICATION RESULTS OF ORIGINAL/PURE FC7 AFTER APPLYING PCA (95%,97%,99%)

Classifiers	Accuracy	Precision	Recall	F measure	Training Time (sec)
fc7 (PCA 95%) results					
ANN	64.977	0.658	0.65	0.648	12295.32
KNN	41.4055	0.509	0.414	0.427	0
3KNN	34.9078	0.502	0.349	0.365	0
5KNN	36.7051	0.52	0.367	0.386	0
Naïve bayes	58.3641	0.643	0.584	0.598	0.06
Random forest	15.8986	0.24	0.159	0.167	15.5
Decision tree	17.0737	0.177	0.171	0.169	40.36
fc7 (PCA 97%) results					
ANN	65.2995	0.66	0.653	0.651	15658
KNN	38.6175	0.532	0.386	0.409	0.01
3KNN	29.7926	0.507	0.298	0.326	0
5KNN	30.4147	0.52	0.304	0.337	0
Naïve bayes	56.9585	0.646	0.57	0.588	0.11
Random forest	12.2811	0.211	0.123	0.13	17.32
Decision tree	16.4977	0.173	0.165	0.164	71.95
fc7 (PCA 99%) results					
ANN	67.9493	0.686	0.679	0.676	23197.76
KNN	27.3272	0.565	0.273	0.324	0.01
3KNN	15.2995	0.45	0.153	0.195	0
5KNN	15.4147	0.464	0.154	0.198	0
Naïve bayes	56.3825	0.678	0.564	0.592	0.53
Random forest	4.3779	0.088	0.044	0.044	22.7
Decision tree	14.7926	0.151	0.148	0.146	137.95

TABLE VI. IDENTIFICATION RESULTS OF AVERAGE BETWEEN (FC6 AND FC7) AFTER APPLYING OF PCA (95%,97%,99%)

Classifiers	Accuracy	Precision	Recall	F measure	Training Time (sec)
AVG (PCA95%) results					
ANN	69.5622	0.703	0.696	0.693	16452.89
KNN	29.1705	0.523	0.292	0.331	0
3KNN	16.6359	0.418	0.166	0.202	0.02
5KNN	16.659	0.429	0.167	0.205	0
Naïve bayes	53.3871	0.635	0.534	0.56	1.35
Random forest	5.7143	0.111	0.057	0.057	15.99
Decision tree	15.2304	0.165	0.152	0.152	85.88
AVG (PCA 97%) results					
ANN	69.9078	0.707	0.699	0.696	24498.83
KNN	20.6221	0.503	0.206	0.254	0
3KNN	8.5484	0.282	0.085	0.109	0
5KNN	8.8249	0.316	0.088	0.115	0
Naïve Bayes	49.7926	0.631	0.498	0.53	0.48
Random forest	3.9862	0.082	0.04	0.039	22.49
Decision tree	14.0553	0.149	0.141	0.14	138.67
AVG (PCA 99%) results					
ANN	65.5069	0.666	0.655	0.652	58379.22
KNN	9.4009	0.34	0.094	0.121	0
3KNN	2.1889	0.089	0.022	0.027	0
5KNN	2.2811	0.094	0.023	0.028	0
Naïve bayes	39.8157	0.607	0.398	0.45	2.2
Random forest	2.0968	0.04	0.021	0.019	26.89
Decision tree	13.9171	0.149	0.139	0.139	314.45

TABLE VII. IDENTIFICATION RESULTS OF MAXIMUM BETWEEN (FC6 AND FC7) AFTER APPLYING OF PCA (95%,97%,99%)

Classifiers	Accuracy	Precision	Recall	F measure	Training Time (sec)
MAX (PCA95%) results					
ANN	66.9816	0.68	0.67	0.668	54151.88
KNN	13.2488	0.432	0.132	0.172	0
3KNN	4.7235	0.169	0.047	0.059	0
5KNN	4.447	0.161	0.044	0.055	0
Naïve bayes	46.5207	0.658	0.465	0.517	1.75
Random forest	2.9032	0.071	0.029	0.032	30.5
Decision tree	15.0922	0.16	0.151	0.15	423.62
MAX (PCA 97%) results					
ANN	-	-	-	-	-
KNN	7.8341	0.32	0.078	0.106	0
3KNN	2.0507	0.07	0.021	0.022	0
5KNN	2.3272	0.073	0.023	0.026	0
Naïve bayes	41.682	0.648	0.417	0.478	2.46
Random forest	2.6959	0.07	0.027	0.028	33
Decision tree	14.8618	0.158	0.149	0.148	556.78
MAX (PCA 99%) results					
ANN	-	-	-	-	-
KNN	3.2258	0.132	0.032	0.043	0
3KNN	0.9908	0.026	0.01	0.008	0
5KNN	1.0599	0.026	0.011	0.009	0
Naïve bayes	32.1889	0.585	0.322	0.38	3.52
Random forest	1.7512	0.043	0.018	0.02	35.19
Decision tree	14.6774	0.156	0.147	0.146	1517.2

Table IX shows the birds identification results in (combine between fc6 and fc7) where the highest accuracy resultant from applying PCA of (95%,97% and 99%) are in favors of ANN with (69.5392, 70.9908 and 67.9263), respectively. The second high accuracy resultant from applying PCA of all percentage of (95%, 97% and 99%) is Naïve Bayes that has achieved accuracy of (57.235, 54.1475 and 43.7558%), respectively.

While for the time spend to conduct the test and training dataset, ANN spend large time (56279.29s). Comparison between the proposal work and previous researchers' works.

Table X compares the results of the proposed approach with three similar approaches for birds identification.

Table X has approved that the output of our proposal can be considered as one of the interesting study compared to the previous researchs, for several reasons:

1) Some of previous studies were conducted on small dataset birds (categories) like in [4], [7] that used (13), (16) categories respectively, compared to this study that used (434).

2) Some others of previous studies conducted on dataset containing a large number of images in training dataset (examples) like in [4], [3], [14] that used (161907), (11788), (11788) examples respectively compared to this study which contained a few images (4340 examples). Few number of images (examples) for each bird usually leads to low accuracy compared to the large examples, but in constant it was not. This leads to make more confident in the results of this study.

3) There were studies conducted for identifying birds using different algorithms and methods based audio/ video like [4][11][6][10], while other studies conducted to identify birds based images using AI algorithms [1][3][17]. This is less in what was conducted in this study that used deep-learning algorithms and different statistical operations like: MAX, MIN, AVERAGE, and combine between the layers fc6/fc7 based on VGG-19 algorithm.

4) This study conducted on different methods like: combine between the fc6/fc7, max of fc6/fc7, min of fc6/fc7, and the average for fc6/fc7 based on VGG-19.

TABLE VIII. IDENTIFICATION RESULTS OF MINIMUM BETWEEN (FC6 AND FC7) AFTER APPLYING OF PCA (95%,97%,99%)

Classifiers	Accuracy	Precision	Recall	F measure	Training Time (sec)
MIN (PCA 95%) results					
ANN	70.8295	0.715	0.708	0.993	42677.02
KNN	17.6037	0.515	0.176	0.223	0
3KNN	6.106	0.234	0.061	0.078	0
5KNN	6.1982	0.238	0.062	0.078	0
Naïve bayes	48.7327	0.661	0.487	0.539	1.2
Random forest	3.5023	0.093	0.035	0.038	33.74
Decision tree	13.6636	0.153	0.142	0.142	2829.83
MIN (PCA 97%) results					
ANN	-	-	-	-	-
KNN	9.5853	0.371	0.096	0.129	0
3KNN	2.5115	0.079	0.025	0.027	0
5KNN	2.6267	0.096	0.026	0.029	0
Naïve bayes	44.1014	0.652	0.441	0.501	2.5
Random forest	2.5115	0.057	0.025	0.025	29.52
Decision tree	13.6636	0.147	0.137	0.136	1007.75
MIN (PCA 99%) results					
ANN	-	-	-	-	-
KNN	3.871	0.176	0.039	0.051	0.01
3KNN	0.8756	0.024	0.009	0.007	0
5KNN	0.7834	0.019	0.008	0.007	0
Naïve bayes	35	0.615	0.35	0.414	3.38
Random forest	2.0507	0.046	0.021	0.021	31.93
Decision tree	13.341	0.144	0.133	0.134	547.58

TABLE IX. IDENTIFICATION RESULTS OF COMBINE ON (FC6 AND FC7) AFTER APPLYING PCA (95%, 97%, 99%)

Classifiers	Accuracy	Precision	Recall	F measure	Training Time (sec)
Combine (PCA 95%) results					
ANN	69.5392	0.703	0.695	0.693	20103.19
KNN	35	0.544	0.35	0.382	0
3KNN	24.3088	0.459	0.243	0.275	0
5KNN	24.1705	0.487	0.242	0.28	0
Naïve bayes	57.235	0.653	0.572	0.592	0.89
Random forest	7.3041	0.166	0.073	0.08	167.7
Decision tree	16.0599	0.167	0.161	0.159	96.93
Combine (PCA 97%) results					
ANN	70.9908	0.718	0.71	0.708	24033.56
KNN	27.1659	0.547	0.272	0.319	0
3KNN	13.7558	0.397	0.138	0.172	0
5KNN	14.5161	0.436	0.145	0.186	0
Naïve bayes	54.1475	0.654	0.541	0.568	0.66
Random forest	5.1152	0.115	0.051	0.054	26.12
Decision tree	15.4839	0.161	0.155	0.153	128.03
Combine (PCA 99%) results					
ANN	67.9263	0.685	0.679	0.675	56279.29
KNN	10.8065	0.39	0.108	0.142	0.02
3KNN	2.9493	0.106	0.029	0.035	0
5KNN	2.9493	0.112	0.029	0.036	0
Naïve bayes	43.7558	0.647	0.438	0.49	3.59
Random forest	2.3733	0.038	0.024	0.022	113.17
Decision tree	14.8618	0.153	0.149	0.147	403.92

TABLE X. COMPARISON BETWEEN PROPOSAL OF THIS STUDY AND RELATED WORKS

	Method	Dataset (name)	# of example	# of category	Result
[4]	CNN+RandomForest	Frames of Videos	161907	13	ACC=90%
[3]	Regularized Softmax Reg w/ Broad Classes	CUB200-2011	11788	200	ACC=70%
[14]	HSV+SVM	CUB200-2011	11788	200	ACC=83.87%
[7]	Mask R-CNN + Ensemble Model	CVIP 2018 Bird Species challenge	150	16	Precision= 56.58
proposal of this study	(Combine original (fc6+fc7) after PCA)+ANN	JOP(new dataset)	4340	434	ACC=71%

VI. CONCLUSION

This study aims at investigating the use of deep learning for birds' identification system using VGG-19 for extracting features from images. VGG-19 is one of the pre-trained convolutional neural network (CNN) networks that used for image identification which was used in this paper to extract the features from birds' images.

Database of this study is contained 4340 images of 434 bird species obtained from scientific sources and where approval by Jordanian Bird Watching Association based on scientific name.

In this study, the two layers in the structure of VGG19 to get the features were used layer 6 (called fc6) and layer 7 (called fc7); each layer consists of 4096 features.

Since the size of the deep feature vector obtained from the VGG19's layers (6 or 7) is very large (4096), we opt for Principal Component Analysis (PCA) and to do the dimensionality reduction. Moreover, it was created more feature vectors called statistical operations to generate more datasets from (fc6 and fc7) using average, minimum, maximum and combine of both layers.

The created datasets (i.e. with PCA and without PCA), as well as the datasets that created from statistical operations are used as input for classification using various machine learning classifiers including Artificial neural networks (ANN), K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes and Decision Tree.

The results of investigation in this study include and not limited to the following, the PCA used on the deep features does not only reduce the dimensionality, and therefore, the

training/testing time is reduced significantly, but also allows for increasing the identification accuracy, particularly when using the ANN classifier. Based on the results of classifiers; ANN showed high classification accuracy (0.9908), precision (0.718), recall (0.71) and F-Measure (0.708) compared to other classifiers.

It is recommended to conduct more investigation to improve accuracy results and to reduce training time using different algorithms.

REFERENCES

- [1] Tayal, Madhuri, Atharva Mangrulkar, Purvashree Waldey, and Chitra Dangra. 2018. "Bird Identification by Image Recognition." *Helix* 8(6): 4349–4352.
- [2] Albustanji, Abeer. 2019. "Veiled-Face Recognition Using Deep Learning." Mutah University.
- [3] Alter, Anne L, and Karen M Wang. 2017. "An Exploration of Computer Vision Techniques for Bird Species Classification."
- [4] Atanbori, John et al. 2018. "Classification of Bird Species from Video Using Appearance and Motion Features" *Ecological Informatics* 48: 12–23.
- [5] Brownlee, Jason. 2016. "How To Use Classification Machine Learning Algorithms in Weka." Retrieved from <https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/>.
- [6] Cai, J., Ee, D., Pham, B., Roe, P., & Zhang, J. (2007, December). Sensor network for the monitoring of ecosystem: Bird species recognition. In 2007 3rd international conference on intelligent sensors, sensor networks and information 293-298. IEEE.
- [7] Kumar, A., & Das, S. D. 2018. "Bird Species Classification Using Transfer Learning with Multistage Training". In *Workshop on Computer Vision Applications* 28-38. Springer, Singapore.
- [8] Hassanat, A. (2018). "Furthest-pair-based binary search tree for speeding big data classification using k-nearest neighbors". *Big Data*, 6(3): 225-235.
- [9] Hijazi, Samer, Rishi Kumar, and Chris Rowen. 2015. "Using Convolutional Neural Networks for Image Recognition." . IP Group, Cadence. Retrieved from https://ip.cadence.com/uploads/901/cnn_wp-pdf.
- [10] Ince, A., Jancsó, H. B., Szilágyi, Z., Farkas, A., & Sulyok, C. 2018. Bird sound recognition using a convolutional neural network. In 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY) :295-300 IEEE.
- [11] Korzh, Oxana, Mikel Joaristi, and Edoardo Serra B. 2018. "Convolutional Neural Network Ensemble Fine-Tuning for Extended Transfer." In *International Conference on Big Data*, 110–23. Retrieved from http://dx.doi.org/10.1007/978-3-319-94301-5_9.
- [12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." 25: 1–9. Technologies.
- [13] The Royal Society For The Conservation Of Nature 2017. "Birdwatching in Jordan". Retrieved from https://migratorysoaringbirds.birdlife.org/sites/default/files/jordan_birding_brochure.pdf.
- [14] Qiao, Baowen, Zuofeng Zhou, Hongtao Yang, and Jianzhong Cao. 2017. "Bird Species Recognition Based on SVM Classifier and Decision Tree." In 2017 First International Conference on Electronics Instrumentation & Information Systems 1–4.
- [15] Sarayrah, Bayan mahmoud. 2019. "Finger Knuckle Print Recognition Using Deep Learning." Mutah University.
- [16] S. Al-Showarah et. al. 2020. "The Effect of Age and Screen Sizes on the Usability of Smartphones Based on Handwriting of English Words on the Touchscreen", *Mu'tah Lil-Buhuth wad-Dirasat, Natural and Applied Sciences series*, Vol. 35, No. 1, 2020. ISSN: 1022-6812.
- [17] Triveni, G., Malleswari, G. N., Sree, K. N. S., & Ramya, M (2020). Bird Species Identification using Deep Fuzzy Neural Network *Int. J. Res. Appl. Sci. Eng. Technol.(IJRASET)*, 8: 1214-1219.
- [18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie 2011. "The Caltech-UCSD Birds-200-2011 Dataset." Technical Report CNS-TR-2011-001, California Institute of Technology.
- [19] Welinder, Peter et al. 2010. "Caltech-UCSD Birds 200." Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Feature Engineering Algorithms for Traffic Dataset

Akibu Mahmoud Abdullah*¹, Raja Sher Afgun Usmani², Thulasyammal Ramiah Pillai³,
Ibrahim Abaker Targio Hashem⁴, Mohsen Marjani⁵
School of Computer Science and Engineering, Taylor's University,^{1,2,3,5}
Selangor, Malaysia
College of Computing and Informatics, Department of Computer Science⁴,
University of Sharjah, 27272 Sharjah, UAE

Abstract—As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. The emergence of an automatic traffic management system allows us to record and monitor motor vehicles' movement in a road segment. One of the challenges researchers face is when the historical traffic data is given as an annual average that contains incomplete data. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am. Hence there is a need to estimate hourly traffic volume for the remaining hours. Feature engineering can be used to overcome the issue of incomplete data. This paper proposed feature engineering algorithms that can efficiently estimate hourly traffic volume and generate features from the existing dataset for all traffic census stations in Malaysia using queuing theory. The proposed feature engineering algorithms were able to estimate the hourly traffic volume and generate features for three years in Jalan Kepong census station, Kuala Lumpur, Malaysia. The algorithms were evaluated using the Random Forest model and Decision Tree Models. The result shows that our feature engineering algorithms improve machine learning algorithms' performance except for the prediction of NO_2 using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction. The algorithm is applied in one of the traffic census stations in Kuala Lumpur, and it can be used for the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes average hourly data.

Keywords—Feature engineering algorithm; queuing theory; Road Traffic Volume Malaysia (RTVM); machine learning algorithms

I. INTRODUCTION

As a result of an increase in the human population globally, traffic congestion in the urban area is becoming worse, which leads to time-consuming, waste of fuel, and, most importantly, the emission of pollutants. Therefore, there is a need to monitor and estimate traffic density. This reason results in the emergence of an automatic traffic management system for recording and monitoring the hourly and daily movement of motor vehicles. Several studies reported that motor vehicles are primary sources of air pollution in the urban area worldwide [1]. The concentration and increase of air pollution depend on the increase of traffic volume, speed of the vehicle, type

of vehicle and many more factors. Researchers are looking for creative solutions like smart cities and GIS systems to avoid traffic congestion and volume [2, 3]. A study conducted reveals that traffic volume has a significant impact on PM_{10} , NO_x , NO , and NO_2 concentrations [4]. [5]'s study shows that the increase of the vehicle increases the concentrations of air pollution during peak hours in the morning and evening. Traffic volume, traffic congestion, and low speed increase level of PM and NO_x emissions [6]. A study in Kuala Lumpur shows that air pollution concentration strongly depends on traffic volume, waiting time on the road, speed of the vehicle, and fuel consumption [7]. Speed of the vehicle, composition, traffic volume, intensity, and acceleration influenced the concentration of air pollution [8].

Number	Year	Time	HTV	Car	Van	Mlorry	Hlorry	Bus	Motorcycl	AWT	Speed
1	2014	0700-0800	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	2014	0800-0900	9415	67%	8.90%	2%	0.20%	0.70%	21.20%	N/A	N/A
3	2014	0900-1000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	2014	1000-1100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	2014	1100-1200	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	2014	1200-1300	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	2014	1300-1400	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	2014	1400-1500	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9	2014	1500-1600	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
10	2014	1600-1700	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
11	2014	1700-1800	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	2014	1800-1900	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
13	2014	1900-2000	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
14	2014	2000-2100	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
15	2014	2100-2200	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
16	2014	2200-2300	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
TOTAL			133,180	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Fig. 1. Summary of Daily and Hourly Traffic Volume.

One of the challenge researchers are facing is when the recorded traffic data is given as an annual average that contains incomplete data- and most of these researchers are conducting multidisciplinary studies- [9]. The annual average daily traffic (AADT) is an average number of traffic volumes at the roadway segment in a specific location over a year. AADT data are collected using surveillance cameras to count and monitor passing vehicles on a 24-hour, 16-hour, 5-hour, or 1-hour basis. These data are used mostly in road transport studies, such as estimation of fuel consumption, roadway planning, emission prediction, traffic operation, travel behavior, accident predictions, and many more [10]. An example of AADT data is the one given by Road Traffic Volume Malaysia (RTVM), and this data is incomplete. The RTVM provides an average of daily traffic data and one peak hour. The recorded traffic data is for sixteen hours, and the only hourly data given is one hour, from 8.00 am to 9.00 am, as shown in Fig. 1. The total daily traffic volume and peak hour (hourly traffic volume from 8.00

to 9.00 am) were highlighted with red color in the Figure. The highlighted blue color indicates the volume of type of the vehicle. The not available (N/A) in the Figure shows that the remaining hourly traffic volume and volume of the type of the vehicles were missing. There is a need to estimate the hourly traffic for the remaining hours. In this study, feature engineering was applied to overcome the issue of incomplete hourly traffic volume.

Feature engineering is one of the most challenging and significant tasks in data science. Extracting and generating new variable from the existing dataset is a difficult task, and also consume time, and effort to process variable in dataset before applying them in the model. Feature engineering is the process of extracting and generating new features or variables from the existing dataset which helps in improving the performance of Machine Learning Algorithms. It also helps to understand the data deeply and gives more valuable insights. Data scientist spend more than 80% of their time on cleaning the dataset [11].

The structure of the paper is presented as follows, Section II presents the related works, Section III discusses the methodology, and it is divided into two sections; namely, Section III-A presented the data and how it was collected, and Section III-B discusses the proposed feature engineering algorithms. In Section IV, the results were presented, it has two sections, Section IV-A is the feature engineering algorithms result and Section IV-B is the prediction of traffic emissions with and without traffic dataset. Discussion was presented in Section V. Lastly, the conclusion is discussed in Section VI.

II. RELATED WORK

Traffic volume is one of the important variable, which contribute for increasing air pollution level produced by automobiles. Many studies were conducted to estimated hourly traffic volume, for example [12] applied extreme gradient boosting tree (XGBoost) and graph theory to estimate hourly traffic volume at location without traffic sensor in Utah United States of America (USA). The developed model was able to estimate hourly traffic volume. Study of [13] propose deep learning algorithm and image processing method to estimate traffic volume, vehicle type, and vehicle speed using recorded traffic video. The model was found good with 90% of accuracy for traffic volume estimation. Estimation of hourly traffic volume was conducted using Artificial Neural Network (ANN) [14]. The applied ANN model was able to estimated hourly traffic volume.

Time spent on the road and speed vehicles were responsible for the variability and trend of air pollution. Several studies were conducted to estimate vehicle speed and time spent on the road. These studies include [15], [16], [17], [18], [19], [20], [21], [22], [23], and [24]. These studies can be divided into three. Firstly, most of the studies have speed parameters in their data, so they developed models to estimate the vehicle speed in a location that they do not have traffic sensor stations. Some researchers proposed an algorithm to estimate the dataset's missing values, while others focus on estimating the speed to evaluate their models using the historical dataset. The second category is having a recorded video of the moving automobiles on the road, so they developed methods to estimate the

vehicle's speed. Lastly, some studies installed sensor devices on the road to calculate and estimate the vehicles' speed.

In general, all these studies used four types of traffic datasets, namely, sensor device data, video-based data, image-based data, and vehicle data. Fig. 2 presents the four types of datasets used for the estimation of vehicle speed and time spent on the road. All of the presented studies none of them estimate or generate vehicle speed and time spent using AADT dataset. To the best of our knowledge, we could not find a study that generates the cars' speed and time spent on the road using AADT dataset.

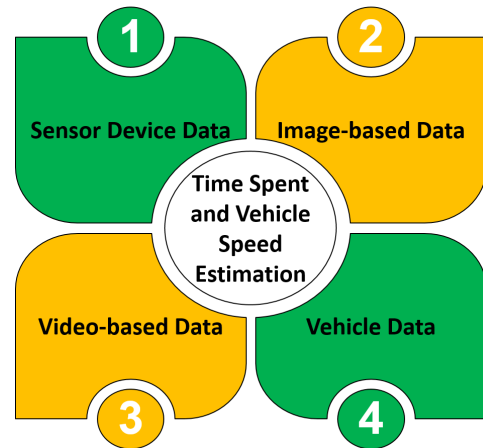


Fig. 2. Types of Dataset Used for Estimation of Vehicle Speed and Time Spent on the Road.

The RTVM data have been used by many studies in Malaysia. Table V presents the researches were conducted in Malaysia using the RTVM dataset. These studies mostly used the data to explore the Level of Service (LOS), the number of registered vehicles, and traffic density. There is a lack of study using hourly traffic data due to unavailability and incomplete hourly traffic volume dataset. In this paper, we proposed feature engineering algorithms which can efficiently estimate hourly traffic volume and generate features from existing traffic dataset (RTVM dataset). Queuing model is proposed to generate vehicle speed and time spent on the road features.

III. METHODOLOGY

A. Data

There is a total number of 554 traffic monitoring stations all over Malaysia. The traffic was recorded hourly for the state and federal roads in Malaysia by the State Public Works Department (JKR Negeri) and organized by Road Traffic Volume Malaysia (RTVM). In 1982, the first printed copy of the organized national traffic census was published by the RTVM. From early 1999, the data were available in Compact Disc (CD). In contrast, the online version started from 2014 to date. The traffic sensor is conducted twice a year during March-April or September-October. The data collection is categorized into three types, type 0, type 1, and type 2. The type 0 data is recorded for 24 hours in 7 days, while type

1 for 16 hours in 7 days, and type 2 for 16 in 1 day. The recording for 16 hours started from 6.00 am to 10.00 pm. The vehicles are divided into six classes, as presented in Table I. We also utilize the air pollution dataset from the AQM stations provided by the Department of Environment (DOE), Malaysia and feature engineered in our previous work [11, 25, 26]

TABLE I. VEHICLE CLASSES

No	Class	Types of Vehicle
1	Class 1	Motor Cars
		Taxi
2	Class 2	Small Vans
		Utilities (Light 2-axles)
3	Class 3	Lorries
		Large Vans (Heavy 2-axles)
4	Class 4	Lorries with 3-axles and above
5	Class 5	Buses
6	Class 6	Motorcycles
		Scooters

The RTVM divides the carriageway into two, single and dual carriageway. Table II describes the carriageway types, their code, and how many lanes each road includes.

TABLE II. TYPE OF CARRIAGEWAY

Code	Single Carriageway	Code	Dual Carriageway
T1-1	Two-Lane	K1+2	Three Lane - One way
T1-2	Three-Lane	K2+2	Four Lane - One way
T2	Two-Lane - One way	K2-2	Four Lane
T2-2	Four-Lane	k(1+2)-(1+2)	Six Lane
T3	Three-Lane - One way	—	—

Fig. 3 summarizes the total number of traffic census stations in each region in Malaysia. Johor is recorded as the highest region with 75 stations, while Kuala Lumpur with the lowest number of stations with 5.

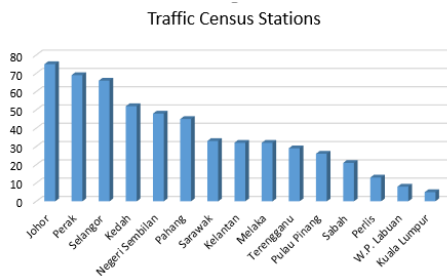


Fig. 3. Traffic Census Stations in Malaysia.

In this study, we proposed a new algorithm for estimating hourly traffic data based on AADT data provided by RTVM. Kuala Lumpur traffic census station was chosen for this study, and it has five stations; these stations are dual carriageways with six lanes. Table III presents the station's ID, locations, and a kilometer of each road. The WR101 station was chosen in this study. The traffic data for 2014, 2015, and 2016 were used in

this study. Table IV summarizes the hourly (one peak hour) and daily (16 hours) data based on the average estimation given by RTVM.

TABLE III. KUALA LUMPUR CENSUS STATION

Station ID	Location	Kilometer
WR101	Kuala Lumpur - Kuala Selangor (Jalan Kepong)	8.9
WR102	Kuala Lumpur - Ipoh	12.1
WR103	Kuala Lumpur Ipoh	8.1
WR105	Kuala Lumpur - Seremban Expressway	8.1
WR106	Kuala Lumpur - Damansara	5.8

TABLE IV. TRAFFIC VOLUME FOR THREE YEARS

Year	Month	Daily	Hourly	Time
2014	March	133180	9415	9-10
	September	131623	10979	8-9
2015	March	119921	9750	10-11
	September	102273	8031	8-9
2016	April	132029	10083	10-11
	October	113347	7787	18-19

B. Feature Engineering Algorithms

1) *Feature Estimation*: The feature engineering algorithm is proposed to estimate hourly traffic volume. The algorithm performs two tasks; EstimationOfData and DataDistribution. The EstimationOfData is performed by selecting the station, peak hour, and the normal hour. Six peak hours (peak hours 7.00 to 8.00 am, 8.00 to 9.00 am, 9.00 to 10.00 am, 10.00 to 11.00 am, 17.00 to 18.00 pm, and 18.00 to 19.00 pm) were selected and distributed randomly from the daily average traffic volume for six months, while the remaining amount of the daily traffic volume were distributed randomly to the ten hours (normal hours 11.00 am to 12.00 pm, 12.00 to 13.00 pm, 13.00 to 14.00 pm, 14.00 to 15.00 pm, 15.00 to 16.00 pm, 16.00 to 17.00 pm, 19.00 to 20.00 pm, 20.00 to 21.00 pm, 21.00 to 22.00 pm, and 22.00 to 23.00 pm) for six months as given in the following equations:

$$p = h \tag{1}$$

$$n = d - p \tag{2}$$

The p is the peak hour, h stand for hourly data given by RTVM, n is the normal hour, d daily traffic volume. The percentage of the vehicle type was distributed from the total daily traffic volume using the below equation. The v is the vehicle type, d is the daily traffic volume, and c is the percentage of vehicle type.

$$v = d * p \tag{3}$$

The DataDistribution is the distribution of the estimated traffic data obtained from EstimationOfData. Since the data is based on a six-month average, we create three-years data with hourly rows for sixteen hours, because the RTVM data is based on sixteen hours. We first distribute the amount of peak

TABLE V. PREVIOUS STUDIES THAT USED RTVM DATASET

Author	Objective	Location	Dataset
[27]	Condition of vehicle engine based on driving behavior	Kuala Lumpur	RTVM_2015
[28]	Effect of traffic-related air pollution on traffic policemen	Klang Valley	RTVM_2017
[29]	Association between traffic-related air pollution with respiratory symptoms and DNA damage	Kajang, Hulu Langat, Selangor	RTVM_2016
[30]	Proposed an intelligent national transportation management center	Kuala Lumpur	RTVM_2016
[31]	The autonomous emergency braking system was proposed for the primary accident that is occurring	Jalan Butterworth, Penang	RTVM_2016
[32]	Proposed preventive model for road maintenance	Petaling	RTVM
[33]	Exploring the effect contributing to the vehicle accident	Sabah	RTVM
[34]	Investigating the influence of rubbernecking towards vehicle deceleration rate due to primary accident in the urban area	Jalan Butterworth and Jalan George Town, Penang	RTVM_2016
[31]	Proposed Malaysia driving cycle (MDC) for light-duty test	Terengganu	RTVM_2015
[35]	Investigates driving cycle which contributes to producing air pollution and fuel consumption	Kuala Lumpur	RTVM
[36]	Estimation of particulate matter from non-exhaust and exhaust vehicle	Klang Valley	RTVM_2014
[34]	Development of the driving cycle for route selection	Penang	RTVM_2015
[37]	Investigates single-vehicle accident along with mountainous areas	Sabah	RTVM_2013
[38]	The concentration of air pollution near a primary schools	Pahang	RTVM_2014
[39]	Driving behavior among heavy truck drivers due to pavement damage	Ampang, Kuala Lumpur	RTVM
[40]	Proposed a model for estimating annual daily traffic for single carriageway	Johor	RTVM_2012
[41]	Level of CO and CO ₂ due to the increase of motor vehicles in industrial areas	Shah Alam, Seremban, and Kuantan	RTVM
[42]	Estimation of CO produced by passenger cars	Selangor	RTVM_2011
[43]	Vehicle compassion in Malaysia and Indonesia	Malaysia	RTVM_2009
[44]	Estimation of future traffic volume	Skudai, Johor	RTVM_2011
[45]	The trend of public transport in Perak	Perak	RTVM_2005

hour for the six hours randomly and then insert and distribute the normal hours randomly (the remaining ten hours), which is the remaining ten hours. Lastly, we distribute the amount to the type of vehicles based on the percentage given in RTVM data. The algorithm is presented in Fig. 4. The six peak hours were distributed randomly using range with minimum and maximum values (so that the values would not be same). Similarly, normal hours were distributed randomly using range with minimum and maximum values.

2) *Feature Generation*: In this study, queuing theory is applied to calculate average speed of vehicle and time spent on the road. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing theory is a mathematical study of estimation of the waiting time in the queue. Queuing system considers arrival time, number of server and service time. The arrival time is considered as the arrival of motor vehicles on the particular road segment, the server is the installed camera that recorded the passing vehicles, while the service rate is the time after the vehicle leaves where the camera was installed. The Queuing model structure is presented in Fig. 5, and the notations used in the model.

The arrival rate of the vehicles is distributed using poison distribution. Similarly, the service rate is exponentially distributed. The time spent or waiting time is calculated using the following equation:

$$w = \frac{\lambda^2}{\mu(\mu - \lambda)} \quad (4)$$

```

Algorithm 1: EstimationOfData&DataDistribution
Input: Daily Traffic Volume
Output: Hourly Traffic Volume Estimation
Data: Peak Hour p
1 foreach Six_month do
2   if p = h then
3     for _ in range(p):
4       value = randit(min, max)
5       print(value)
6     /* The p is the peak hour, h stand for hourly data given by RTVM.
7       min is minimum and max is the maximum */
8   else if != p then
9     n = d - p
10    for _ in range(n):
11      value = randit(min, max)
12      print(value)
13    /* The n is the normal hour, d daily traffic volume */
14  foreach type_of_vehicle do
15    if v = c then
16      /* The v is the type of vehicle, d is the daily traffic volume,
17        and c is the percentage of vehicle type */
18      v ++
19    else if != v then
20      END
21 END

```

Fig. 4. Feature Estimation Algorithm.

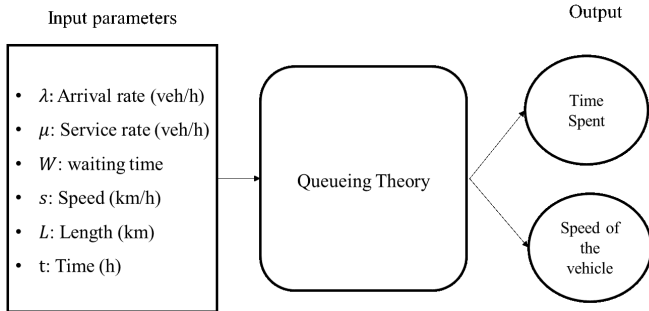


Fig. 5. Queuing Theory & Notations.

The average speed of the motor vehicles is calculated using the below equation:

$$s = \frac{L}{t} \quad (5)$$

IV. RESULT

A. Feature Engineering Algorithms

Motor vehicle has a significant impact on traffic emissions concentration, and human health; moreover, it causes accidents, traffic congestion, and fuel consumption. The emergence of an automatic traffic management system allows us to record and monitor every vehicle passing on the road. The recorded data are used primarily in transportation studies. In Malaysia, the RTVM provided annual average traffic data. There is a need to estimate the hourly traffic volume. This study proposed a feature engineering algorithm that will efficiently estimate hourly traffic volume and generate features from the existing dataset in Malaysia’s traffic census stations. The RTVM gives an average hourly (for peak hour) and daily traffic volume in specific stations. The proposed algorithm was able to estimate the hourly traffic volume for three years based on the yearly average provided by RTVM and distributed the types of vehicles based on the percentage given in the data. Furthermore, the queuing theory was able to generate the vehicle’s average speed and time spent on the road. The output of the feature engineering algorithms (estimated and generated features) was shown in Fig. 6.

Number	Year	Time	HTV	Cars	Van	MLorries	HLorries	Bus	Motorcycle	WT	Speed
1	2014	0700-0800	9662	6476	890	193	19	68	2048	18	74.16
2	2014	0800-0900	10017	6772	900	202	20	71	2148	18	74.16
3	2014	0900-1000	9430	6318	839	189	19	66	1995	18	74.16
4	2014	1000-1100	9589	6425	853	192	19	67	2033	17	74.16
5	2014	1100-1200	7074	4740	630	141	14	50	1500	13	74.16
6	2014	1200-1300	7519	5038	669	150	15	53	1594	13	74.16
7	2014	1300-1400	7804	5229	695	156	16	55	1654	15	74.16
8	2014	1400-1500	7758	5158	690	155	16	54	1643	11	74.16
9	2014	1500-1600	7719	5172	687	154	15	54	1636	15	74.16
10	2014	1600-1700	7001	4691	623	140	14	49	1484	14	74.16
11	2014	1700-1800	10307	6906	917	206	21	72	2182	20	74.16
12	2014	1800-1900	9786	6557	871	196	20	69	2072	18	74.16
13	2014	1900-2000	7198	4823	641	144	14	50	1526	15	74.16
14	2014	2000-2100	7057	4728	628	141	14	49	1496	15	74.16
15	2014	2100-2200	7444	4987	663	149	15	52	1576	15	74.16
16	2014	2200-2300	7716	5170	687	154	15	54	1638	14	74.16

Fig. 6. Feature Engineering Algorithms Result.

Fig. 6 shows the estimated features from the left, which was highlighted with red color. The estimated features were Hourly Traffic Volume (HTV), and types of vehicles (Car and Taxi, Van and Utilities, Medium Lorries, Heavy Lorries, Buses, and

Motorcycles) and generated feature from the right with blue color highlighted (Waiting time on the road and average speed of the vehicle).

B. Prediction of Traffic Emission With and Without Traffic Dataset

Due to the lack of studies that generate and estimate features from the RTVM dataset. To justify the claim that feature engineering improves machine learning models’ performance, we proposed Random Forest and Decision Three machine learning algorithms to predict traffic emissions concentrations using the estimated and generated features (traffic dataset). Additional dataset of air quality and meteorological variables in [11] study were used. The input and output variables were presented in Table VI.

TABLE VI. INPUT & OUTPUT FEATURES

Input Variables		Output Variables
Traffic Variables	Meteorological Variables	Air Pollutants
Traffic Volume, Types of vehicle (Car and Taxi, Van and Utilities, Medium Lorries, Heavy Lorries, Buses, and Motorcycles.	Wind Speed, Wind Direction, Temperature, and Relative Humidity.	Carbon Monoxide (CO), Nitrogen Monoxide(NO), Nitrogen Dioxide (NO ₂), Nitrogen Oxides (NO _x).

Evaluation metrics such as Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) were used to evaluate the performance of the models. First of all, we predict the level of the CO, NO, NO₂, and NO_x pollutants using meteorological features but without traffic data. Lastly, we included the traffic dataset (estimated and generated features) for prediction. The result shows that our feature engineering algorithms improve the accuracy of the machine learning models (Random Forest and Decision Tree models) by predicting the level of traffic pollutants except for the NO₂, which shows no improvement using the Random Forest model as presented in Table VII and VIII. We can also visualize the results in Figures 7, 8, and 9 for the Random Forest Model, while Figures 10, 11, and 12 for the Decision Tree Model.

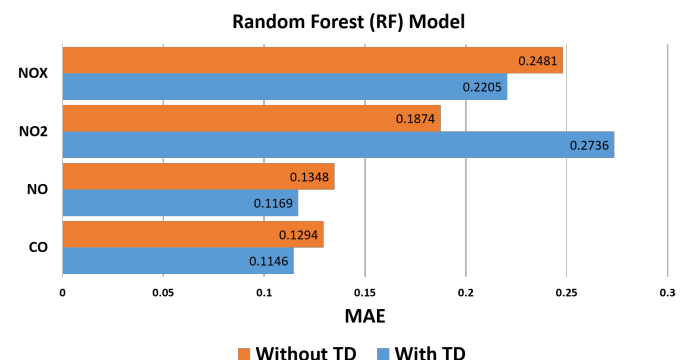


Fig. 7. MAE for the Random Forest Model.

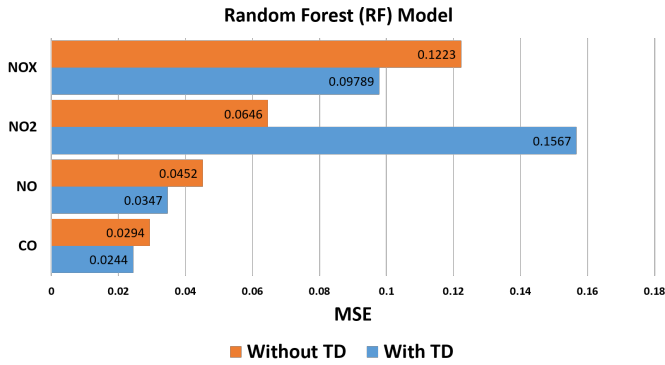


Fig. 8. MSE for the Random Forest Model.

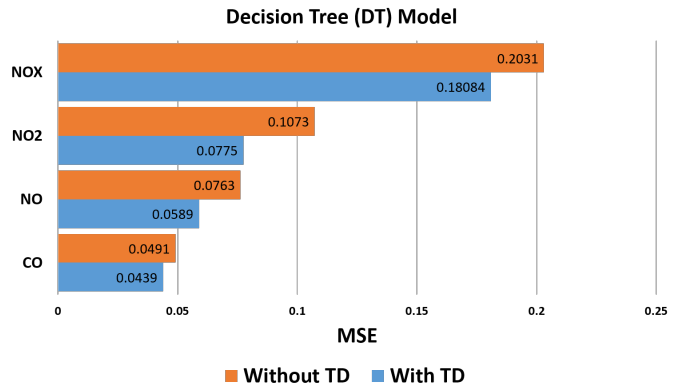


Fig. 11. MSE for the Decision Tree Model.

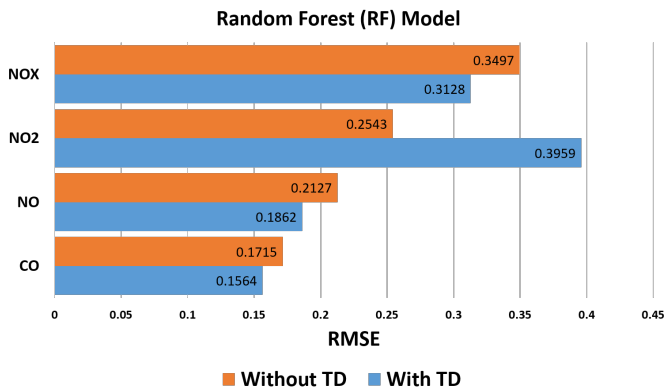


Fig. 9. RMSE for the Random Forest Model.

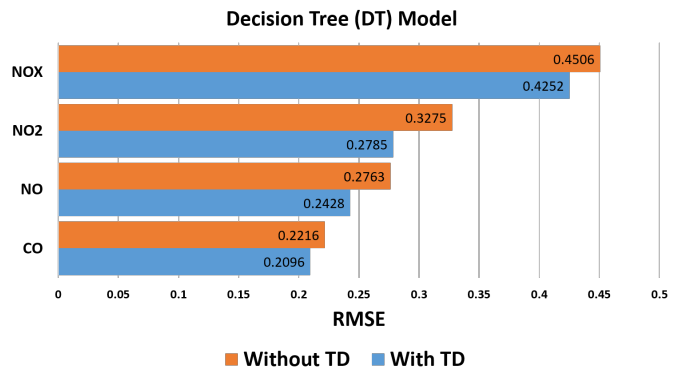


Fig. 12. RMSE for the Decision Tree Model.

TABLE VII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING RANDOM FOREST MODEL

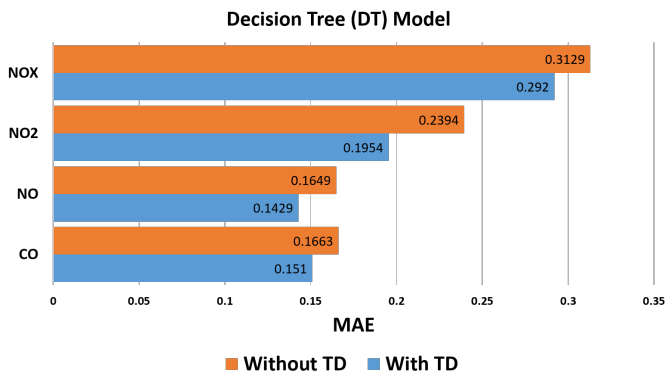
Pollutant	Random Forest (RF) Model					
	Without Traffic Data			With Traffic Data		
	MAE	MSE	RMSE	MAE	MSE	RMSE
CO	0.1294	0.0294	0.1715	0.1146	0.0244	0.1564
NO	0.1348	0.0452	0.2127	0.1169	0.0347	0.1862
NO ₂	0.1874	0.0646	0.2543	0.2736	0.1567	0.3959
NO _x	0.2481	0.1223	0.3497	0.2205	0.09789	0.3128

TABLE VIII. PREDICTION WITH AND WITHOUT TRAFFIC DATA USING DECISION TREE MODEL

Pollutant	Decision Tree (DT) Model					
	Without Traffic Data			With Traffic Data		
	MAE	MSE	RMSE	MAE	MSE	RMSE
CO	0.1663	0.0491	0.2216	0.151	0.0439	0.2096
NO	0.1649	0.0763	0.2763	0.1429	0.0589	0.2428
NO ₂	0.2394	0.1073	0.3275	0.1954	0.0775	0.2785
NO _x	0.3129	0.2031	0.4506	0.292	0.18084	0.4252

Fig. 7, 10, 8, 11, 9, and 12 show that the performance of RF model for prediction with traffic data were better than the DT with lower MAE, MSE, and RMSE for CO, NO, and NO_x, except NO₂ which show higher MAE, MSE, and RMSE. Similarly, for the prediction without traffic dataset, the RF performed better than DT with lower MAE, MSE and

Fig. 10. MAE for the Decision Tree Model.



RMSE for CO , NO , NO_2 and NO_x . We can conclude that the Random Forest Model was a good choice for the prediction of traffic emission concentrations using traffic dataset and without traffic dataset also.

V. DISCUSSION

Motor vehicles become one of the primary concern for the government and agencies. Automobiles create many problems such accident and emissions of pollution to atmosphere. The air pollution produced by motor vehicles had significant impact on human health and the environment as well. Several studies have been conducted for estimation and prediction of traffic emissions. The variability and increase of air pollution depend on traffic characteristics. One of the issue researchers are facing is when the traffic data was provided as an annual average daily traffic (ADDT). The RTVM provides ADDT dataset. The data were incomplete as shown in the Fig. 1. We proposed feature engineering algorithms to estimated and generate missing values in the RTVM dataset. Our feature engineering algorithms were able to generate and estimate the missing features as presented in the Fig. 6. Our feature engineering algorithm is applied in one of the stations in Kuala Lumpur traffic census stations, and this algorithm can be used on the other stations in Malaysia. Additionally, the algorithm can also be used for any annual average daily traffic data if it includes an average hourly dataset. There some limitations in this study, firstly, the estimated hourly traffic volume is proposed due to insufficient hourly traffic volume, which may not provide the exact traffic volume hourly. The RTVM does not provide speed of the vehicle, we calculate vehicle speed as an average basis (the speed of the vehicle is constant). This study could not extract acceleration/deceleration. Some studies suggested that different types of fuel have different emissions, but consideration of fuel type is not provided in this study. Jalan Kepong traffic census station was the selected station in this study, the remaining stations were not studied.

VI. CONCLUSION

Motor vehicles are the primary source of air pollution in metropolitan globally. Air pollution has a significant effect on human health with diseases such as asthma, cardiovascular, and respiratory. Motor vehicle also causes accidents and create congestion at road segments. Due to these reasons, the government and agencies introduce an automated traffic management system to record the passing vehicles on the road. The recorded data has been used in various studies by researchers. The Road Traffic Volume Malaysia provides incomplete traffic data. We proposed a new feature engineering algorithm to overcome the issue of incomplete traffic data. The proposed feature engineering algorithms could estimate the hourly traffic volume and generate features for three years in Jalan Kepong, Kuala Lumpur, Malaysia. The algorithm was evaluated by predicting four traffic pollutants CO , NO , NO_2 , and NO_x using Random Forest and Decision Tree models. The prediction was conducted in two phases, phase one is prediction without traffic dataset (estimated and generated features), and phase two is the prediction with traffic dataset. The result shows that our feature engineering algorithms improve machine learning models' performance except for the prediction of NO_2 using Random Forest, which shows the highest MAE, MSE, and RMSE when traffic data was included for prediction.

ACKNOWLEDGMENT

This research is funded by Taylor's University under the research grant application ID (TUFR/2017/004/04) entitled as "Modeling and Visualization of Air-Pollution and its Impacts on Health". We are also thankful to Department of Environment, Malaysia for providing the AQM station dataset and Ministry of Works, Malaysia for providing the RTVM dataset.

REFERENCES

- [1] R. S. A. Usmani, A. Saeed, A. M. Abdullahi, T. R. Pillai, N. Z. Jhanjhi, and I. A. T. Hashem, "Air pollution and its health impacts in Malaysia: a review," *Air Quality, Atmosphere & Health*, jul 2020. [Online]. Available: [https://doi.org/10.1007/s11869-020-00867-x](https://doi.org/10.1007/s11869-020-00867-xhttp://link.springer.com/10.1007/s11869-020-00867-x)
- [2] R. S. A. Usmani, I. A. T. Hashem, T. R. Pillai, A. Saeed, and A. M. Abdullahi, "Geographic Information System and Big Spatial Data," *International Journal of Enterprise Information Systems (IJEIS)*, vol. 16, no. 4, 2020.
- [3] M. Bilal, R. S. A. Usmani, M. Tayyab, A. A. Mahmoud, R. M. Abdalla, M. Marjani, T. R. Pillai, and I. A. Targio Hashem, "Smart Cities Data: Framework, Applications, and Challenges BT - Handbook of Smart Cities," J. C. Augusto, Ed. Cham: Springer International Publishing, 2020, pp. 1–29. [Online]. Available: https://doi.org/10.1007/978-3-030-15145-4{_}6-1
- [4] R. Rossi, R. Ceccato, and M. Gastaldi, "Effect of road traffic on air pollution. experimental evidence from covid-19 lockdown," *Sustainability*, vol. 12, no. 21, p. 8984, 2020.
- [5] P. Krecel, Y. A. Cipoli, A. C. Targino, L. B. Castro, L. Gidhagen, F. Malucelli, and A. Wolf, "Cyclists' exposure to air pollution under different traffic management strategies," *Science of the Total Environment*, vol. 723, p. 138043, 2020.
- [6] N. Abdull, M. Yoneda, and Y. Shimada, "Traffic characteristics and pollutant emission from road transport in urban area," *Air Quality, Atmosphere & Health*, vol. 13, no. 6, pp. 731–738, 2020.
- [7] S. S. Anjum, R. M. Noor, N. Aghamohammadi, I. Ahmedy, L. M. Kiah, N. Hussin, M. H. Anisi, and M. A. Qureshi, "Modeling traffic congestion based on air quality for greener environment: an empirical study," *IEEE Access*, vol. 7, pp. 57 100–57 119, 2019.
- [8] O. V. Kurnykina, O. V. Popova, S. V. Zubkova, D. V. Karpukhin, V. P. Pavlov, P. K. Varenik, I. A. Aleshkova, and L. Y. Novitskaya, "Air pollution by road traffic and its measurement methods," *EurAsian Journal of BioSciences*, vol. 12, no. 2, pp. 181–188, 2018.
- [9] V. Kumar, J. Prasad, and B. Singh, "Traffic density estimation using progressive neural architecture search," *Journal of Statistics and Management Systems*, vol. 23, no. 2, pp. 481–493, 2020.
- [10] A. Sfyridis and P. Agnolucci, "Annual average daily traffic estimation in england and wales: An application of clustering and regression modelling," *Journal of Transport Geography*, vol. 83, p. 102658, 2020.
- [11] R. S. A. Usmani, W. N. F. B. W. Azmi, A. M. Abdullahi, I. A. T. Hashem, and T. R. Pillai, "A novel feature engineering algorithm for air quality datasets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 3, sep 2020.
- [12] Z. Yi, X. C. Liu, N. Markovic, and J. Phillips, "Inferencing hourly traffic volume using data-driven machine learning and graph theory," *Computers, Environment and Urban Systems*, vol. 85, p. 101548, 2021.
- [13] Z. Dai, H. Song, H. Liang, F. Wu, X. Wang, J. Jia, and Y. Fang, "Traffic parameter estimation and control system based on machine vision," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2020.
- [14] S. Zahedian, P. Sekuła, A. Nohekhan, and Z. Vander Laan, "Estimating hourly traffic volumes using artificial neural network with additional inputs from automatic traffic recorders," *Transportation Research Record*, vol. 2674, no. 3, pp. 272–282, 2020.
- [15] C. Zhang, S. Shen, H. Huang, and L. Wang, "Estimation of the vehicle speed using cross-correlation algorithms and mems wireless sensors," *Sensors*, vol. 21, no. 5, p. 1721, 2021.
- [16] C.-H. Chen, "A cell probe-based method for vehicle speed estimation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 103, no. 1, pp. 265–267, 2020.
- [17] Z. Zhang and X. Yang, "Freeway traffic speed estimation by regression machine-learning techniques using probe vehicle and sensor detector

- data," *Journal of transportation engineering, Part A: Systems*, vol. 146, no. 12, p. 04020138, 2020.
- [18] L. R. Costa, M. S. Rauen, and A. B. Fronza, "Car speed estimation based on image scale factor," *Forensic science international*, vol. 310, p. 110229, 2020.
- [19] Y. Feng, G. Mao, B. Cheng, C. Li, Y. Hui, Z. Xu, and J. Chen, "Mag-monitor: Vehicle speed estimation and vehicle classification through a magnetic sensor," *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [20] G. Sil, S. Nama, A. Maji, and A. K. Maurya, "Effect of horizontal curve geometry on vehicle speed distribution: a four-lane divided highway study," *Transportation letters*, vol. 12, no. 10, pp. 713–722, 2020.
- [21] C.-H. Yang and H.-M. Tsai, "Vehicle counting and speed estimation with rfid backscatter signal," in *2019 IEEE Vehicular Networking Conference (VNC)*. IEEE, 2019, pp. 1–8.
- [22] F. Afifah, S. Nasrin, and A. Mukit, "Vehicle speed estimation using image processing," *Journal of Advanced Research in Applied Mechanics*, vol. 48, no. 1, pp. 9–16, 2018.
- [23] H. Dong, M. Wen, and Z. Yang, "Vehicle speed estimation based on 3d convnets and non-local blocks," *Future Internet*, vol. 11, no. 6, p. 123, 2019.
- [24] A. Kumar, P. Khorramshahi, W.-A. Lin, P. Dhar, J.-C. Chen, and R. Chellappa, "A semi-automatic 2d solution for vehicle speed estimation from monocular videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 137–144.
- [25] R. S. A. Usmani, T. R. Pillai, I. A. T. Hashem, N. Z. Jhanjhi, and A. Saeed, "A Spatial Feature Engineering Algorithm for Creating Air Pollution Health Datasets," 2020. [Online]. Available: https://www.techrxiv.org/articles/preprint/A_{__}Spatial_{__}Feature_{__}Engineering_{__}Algorithm_{__}for_{__}Creating_{__}Air_{__}Pollution_{__}Health_{__}Datasets/12376427/2
- [26] —, "A Spatial Feature Engineering Algorithm for Creating Air Pollution Health Datasets," nov 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2666307420300115https://www.techrxiv.org/articles/preprint/A_{__}Spatial_{__}Feature_{__}Engineering_{__}Algorithm_{__}for_{__}Creating_{__}Air_{__}Pollution_{__}Health_{__}Datasets/12376427/2
- [27] M. F. M. Suhaimi, N. A. M. Salleh, M. S. Sarip *et al.*, "Development of kuala lumpur driving cycle for the estimation of fuel consumption and vehicular emission," in *IOP Conference Series: Materials Science and Engineering*, vol. 834, no. 1. IOP Publishing, 2020, p. 012040.
- [28] N. F. M. Fandi, J. Jalaludin, M. T. Latif, H. H. Abd Hamid, M. F. Awang *et al.*, "Btex exposure assessment and inhalation health risks to traffic policemen in the klang valley region, malaysia," *Aerosol and Air Quality Research*, vol. 20, no. 9, pp. 1922–1937, 2020.
- [29] I. N. Ismail, J. Jalaludin, S. A. Bakar, N. H. Hisamuddin, and N. F. Suhaimi, "Association of traffic-related air pollution (trap) with dna damage and respiratory health symptoms among primary school children in selangor," *Asian Journal of Atmospheric Environment (AJAE)*, vol. 13, no. 2, 2019.
- [30] N. S. Musa, N. M. M. Noor, and J. M. Marjan, "The benefits of national intelligent transportation management centre (nitmc) establishment in malaysia," in *IOP Conference Series: Materials Science and Engineering*, vol. 512, no. 1. IOP Publishing, 2019, p. 012013.
- [31] M. Rani, A. Mahayadin, A. Shahrman, Z. Razlan, I. Zunaidi, W. Wan, A. Harun, M. Hashim, I. Ibrahim, N. Kamarrudin *et al.*, "Terengganu routes representation for development of malaysia driving cycle: Route selection methodology," in *IOP Conference Series: Materials Science and Engineering*, vol. 429, no. 1. IOP Publishing, 2018, p. 012050.
- [32] R. Hamsan, H. Hafiz, A. Azlan, M. Keprawi, A. Malik, A. Adamuddin, A. Abdullah, and A. Shafie, "Pavement condition assessment to forecast maintenance program on jkr state roads in petaling district," in *AIP Conference Proceedings*, vol. 1930, no. 1. AIP Publishing LLC, 2018, p. 020021.
- [33] R. Rusli, M. M. Haque, A. P. Afghari, and M. King, "Applying a random parameters negative binomial lindley model to examine multi-vehicle crashes along rural mountainous highways in malaysia," *Accident Analysis & Prevention*, vol. 119, pp. 80–90, 2018.
- [34] A. Mahayadin, A. Shahrman, M. Hashim, Z. Razlan, M. Faizi, A. Harun, N. Kamarrudin, I. Ibrahim, M. Saad, M. Rani *et al.*, "Efficient methodology of route selection for driving cycle development," in *Journal of Physics: Conference Series*, vol. 908, no. 1. IOP Publishing, 2017, p. 012082.
- [35] R. Tharvin, N. Kamarrudin, A. Shahrman, I. Zunaidi, Z. Razlan, W. Wan, A. Harun, M. Hashim, I. Ibrahim, M. Faizi *et al.*, "Development of driving cycle for passenger car under real world driving conditions in kuala lumpur, malaysia," in *IOP Conference Series: Materials Science and Engineering*, vol. 429, no. 1. IOP Publishing, 2018, p. 012047.
- [36] R. E. Elhadi, A. M. Abdullah, A. H. Abdullah, Z. H. Ash'aari, D. Gumel, M. A. Jamalani, L. K. Chng, and F. M. Binyehmed, "A gis-based emission inventory at 1 km-1km spatial resolution for particulate matter (pm10) in klang valley, malaysia," *Science International*, vol. 29, no. 2, pp. 49–49, 2017.
- [37] R. Rusli, M. M. Haque, M. King, and W. S. Voon, "Single-vehicle crashes along rural mountainous highways in malaysia: an application of random parameters negative binomial model," *Accident Analysis & Prevention*, vol. 102, pp. 153–164, 2017.
- [38] M. Zahaba, H. Abdul Hadi, H. Ariffin, N. Abdull, N. Samsuddin, and M. S. Mohd Aris, "Composition and source determination of heavy metals (hm) in particles in selected primary schools in pahang," *ESTEEM Academic Journal*, vol. 13, pp. 176–184, 2017.
- [39] N. A. Khalifa, A. Alnose, A. Zulkiple, and R. Z. Abidin, "Non-pragmatic data collection for road pavement damage on access road to residential estate and the statistical analysis choice," *International Journal of Traffic and Transportation Engineering*, vol. 5, pp. 83–90, 2016.
- [40] N. S. M. Nor, O. C. Puan, N. Mashros, and M. K. B. Ibrahim, "Estimating average daily traffic using alternative method for single carriageway road in southern region malaysia," *ARP Journal of Engineering and Applied Sciences*, 2006.
- [41] A. AZHARI, A. F. MOHAMED, and M. T. LATIF, "Carbon emission from vehicular source in selected industrial areas in malaysia," *International Journal of the Malay World and Civilisation*, vol. 4, no. 1, pp. 89–93, 2016.
- [42] R. E. Elhadi, D. Y. Gumel, and M. Fallah, "Co2 emission inventory of onroad vehicles in selangor state inpeninsular malaysia," *International Journal of Advanced Scientific and Technical Research*, 2015.
- [43] L. S. Putranto, J. Prasetijo, and N. L. P. S. E. Setyarini, "Vehicle composition in indonesia and malaysia," *The 10th Eastern Asia Society for Transportation Studies*, 2013.
- [44] A. Minhans, N. H. Zaki, and R. Belwal, "Traffic impact assessment: A case of proposed hypermarket in skudai town of malaysia," *Jurnal Teknologi*, vol. 65, no. 3, 2013.
- [45] W. Suwardo, M. Napiah, and I. Kamaruddin, "Review on motorization and use of public transport in perak malaysia: realities and challenges," *2nd INTERNATIONAL CONFERENCE ON BUILT ENVIRONMENT IN DEVELOPING COUNTRIES*, 2008.

PlexNet: An Ensemble of Deep Neural Networks for Biometric Template Protection

Ashutosh Singh¹
Institute of Engineering
and Technology,
Dr. A.P.J. Abdul Kalam Technical
University, Uttar Pradesh,
Lucknow, India

Ranjeet Srivastva²
Babu Banarasi Das Northern
India Institute of Technology,
Dr. A.P.J. Abdul Kalam Technical
University, Uttar Pradesh,
Lucknow, India

Yogendra Narain Singh³
Institute of Engineering
and Technology,
Dr. A.P.J. Abdul Kalam Technical
University, Uttar Pradesh,
Lucknow, India

Abstract—The security of biometric systems, especially protecting the templates stored in the gallery database, is a primary concern for researchers. This paper presents a novel framework using an ensemble of deep neural networks to protect biometric features stored as a template. The proposed ensemble chooses two state-of-the-art CNN architectures i.e., ResNet and DenseNet as base models for training. While training, the pre-trained weights enable the learning algorithm to converge faster. The weights obtained through the base model is further used to train other compatible models, generating a fine-tuned model. Thus, four fine-tuned models are prepared, and their learning are fused to form an ensemble named as PlexNet. To analyze biometric templates' security, the rigorous learning of ensemble is collected using a smart box i.e., application programming interface (API). The API is robust and correctly identifies the query image without referring to a template database. Thus, the proposed framework excludes the templates from database and performed predictions based on learning that is irrevocable.

Keywords—Biometrics; template protection; deep learning; transfer learning; ensemble

I. INTRODUCTION

Identity theft is one of the major epidemics of current century. In the absence of a reliable identity proofing system, data and information thefts have plagued the applications such as online transactions and social welfare schemes [1]. Traditional security systems such as ID cards and passwords cannot protect from digital impersonation thus, obsolete due to their lost and stolen possibilities [2]. To overcome the problems of lost and deliberate sharing of identity markers, biometric recognition has introduced and gradually became the prime tool for individual authentication. Biometrics uses the physiological or behavioural traits of an individual for identification. Commonly used physiological traits include face, fingerprint, palmprint and hand geometry, while signatures, gait, voice are used as behavioural traits. These biometric traits are proved to be personal, reliable, accessible and universal [3].

Although, the biometric systems are reliable in comparison to traditional identification systems, they are vulnerable to several exploits. The biometric system vulnerabilities can be broadly categorized as faults, failures and attacks [4]. Faults are mistakes made by human or system such as data corruption, software aging and storage space fragmentation [5]. A failure is the consequence of one or more faults. It may be service failures, development failures or security failures. Unexpected

service, wrong prediction of complexity or unaccounted situations and imbalance thresholds are some examples of service, development and security failures, respectively.

The faults and failures may lead to fraudulent attacks on machine (i.e., hardware and software) as well as at the administrative level [2]. These attacks can be broadly classified as the sensor level, feature extraction level, matcher level and database level attacks. At sensor level, covertly acquired fake samples such as digital face images, synthetic fingerprints or recorded voice can be presented. The replay of raw biometric trait or injection of false data before pre-processing or feature extraction are common feature extractor level attacks. The attacks on communication channel between feature extraction and matcher module is considered as matcher level attack. For example, infested bug to the algorithm or alteration of match score are some matcher level attacks [6].

The attacks on database consists of templates, are possibly the most prominent fraudulent attacks on a biometric system. The template includes biometric characteristics of an individual that may be compromised, if attacked. Storage of templates in diverse applications create a serious threat to the user's privacy [27]. Thus, a robust mechanism is required to secure the templates stored in the database. An ideal mechanism of template protection should meet the following requirements [8]-[10], [29].

- **Cancelability:** The compromised template must be revoked by reissuing a unique template from the same biometric features.
- **Diversity:** It defends user privacy by ensuring that same template is not being used across databases.
- **Security:** The recovery of original template from the compromised ones must be computationally harder. It avoids the fabrication of a physical spoof from the compromised template.
- **Performance:** The template protection mechanism should not affect the recognition performance of the biometric system.

Despite several template protection schemes, the protection of biometric templates while preserving its discriminability is still a challenge [8]-[9], [12]-[28], [30], [16]-[17]. In this paper, a novel framework is designed to collectively meet all

the requirements of a robust biometric template method. It achieves high performance through rigorous learning of ensemble that forms the basis for exclusion of template from the database. The absence of a template ensures the cancelability and security requirements, whereas immovable learning avoids its use in diverse applications.

The proposed ensemble is named 'PlexNet' due to its resemblance to the plexus such as network of nerves in the nervous system [7]. It is a network of pre-trained and fine-tuned architectures of the variants of deep networks *i.e.*, ResNet and DensNet thus, creating four network paths as shown in Fig. 1. Initially at each path, one of the architectures from ResNet or DenseNet is used as base model trained on biometric sample *e.g.*, facial images with weights of ImageNet database. The weights obtained from base model are further fine-tuned with a compatible architecture. The combination of fine-tuned weights from all network paths results an ensemble the 'PlexNet'.

The contribution of the paper is for the protection of biometric templates using ensemble learning. The plexnet is prepared using ensemble of transfer learning. As transfer learning is a well-established method for converging to the desired goal faster and more accurate, hence we are implementing transfer learning as a tool. Also, the existing ensemble methods generally use majority voting or averaging mechanism that made predictions based on the individual training of the models. Whereas, the proposed ensemble method combines the learning by fine stacking all the fine-tuned models and then re-trains on the target dataset. Therefore, the training of ensemble takes the advantage of the weights gathered by each of the fine-tuned model and converges faster with better accuracy than the traditional ensemble models.

The learning of the PlexNet is conducted in an application programming interface (API). The API acts as an smart box to predict the class label for the presented biometric template *i.e.*, facial image as shown in Fig. 2. The learning at API is done so vigorous that it correctly identifies the input image without referring to a template. Thus, the prepared API eliminates the requirement of nurturing a template database. Further, it follows the requirements of a robust template protection method. For example, cancelability *i.e.*, there is no chance to compromise the templates. The learning of the model can not be directly transferred to other models, thus, diversity is ensured. It matches security paradigms, as the recovery of original image is almost impossible. The high performance reported by our proposed PlexNet framework forms the basis for exclusion of the template database. Briefly, the substantial contributions of this work are as follows:

1. A novel ensemble of deep neural networks is designed for biometric template security, as called 'PlexNet' due to its resemblance to the plexus in the nervous system.
2. An innovative learning procedure is adopted using pre-trained models that results in faster training and testing, that further solves the problem of class imbalance.
3. Primally, application program interface (API) is constructed that predicts the output based on rigorous learning without referring to a template database.

4. The efficacy of proposed framework is evaluated on challenging databases of face biometrics *e.g.*, VG-Face2 and MegaFace. PlexNet outperforms other methods and supports presented method of template exclusion thus, making it impregnable.

The rest of the paper is organized as follows. The literature survey of the biometric template protection is presented in Section II. The proposed ensemble framework with a detail description of architectures used and the requirement of transfer learning, ensemble learning and data augmentation are discussed in Section III. The experimentation process that includes database description, experimental setup, results along with security analysis are presented in Section IV. Finally, the conclusions are drawn in Section V.

II. LITERATURE SURVEY AT A GLANCE

The rapid growth in technology and sensing have increased the demand of deployment of the biometric recognition system *e.g.*, at residences, offices, portable devices and other public access checkpoints. Though, this easiness has brought numerous challenges of biometric vulnerabilities that effects the system may be exploited by an adversary. Thus, an efficient and robust mechanism for biometric template protection is essential. The information captured by the sensor may differ for the same biometric trait. So, the conventional method of password protection using cryptographic hash functions can not be directly applied to a biometric system. Normal encryption is not a sustainable solution, as the saved template must be decrypted every time to make the comparison with the query template. Several biometric template protection techniques have been developed [8], [9], [12]-[28], [30], [16]-[17]. Various biometric template protection techniques are classified as follows,

1. *Biometric Cryptosystems*: Biometric cryptosystem was originally developed to protect keys either using biometric features or generating a cryptographic key [8], [9]. Later, it evolved as a technique for template protection. In a biometric cryptosystem, public information known as helper data is stored in the database instead of original biometric template [10]. This helper data is independent of biometric template and do not reveal the biometric features. It just helps to recover the key used for encryption. Hence, matching is performed between the key extracted from the query template and the stored template key. Biometric cryptosystems can be classified into key binding and key generation approaches [48]. These algorithms differ in the way they generate the helper data.
 - (a) *Key binding approach*: In the key binding approach, the helper data is obtained by associating a key with the biometric template. The helper data is stored as a template in the database. Since the key is independent of the biometric data, it is really hard to recover the original template or the binding key for the helper data. The stored template is used for key recovery at the time of matching with query template. Thus, the authentication is performed with matching of query and stored keys. This approach is primarily used for key protection. The utilization of these approaches for template protection may not fulfil the criteria of cancelability. Here, the

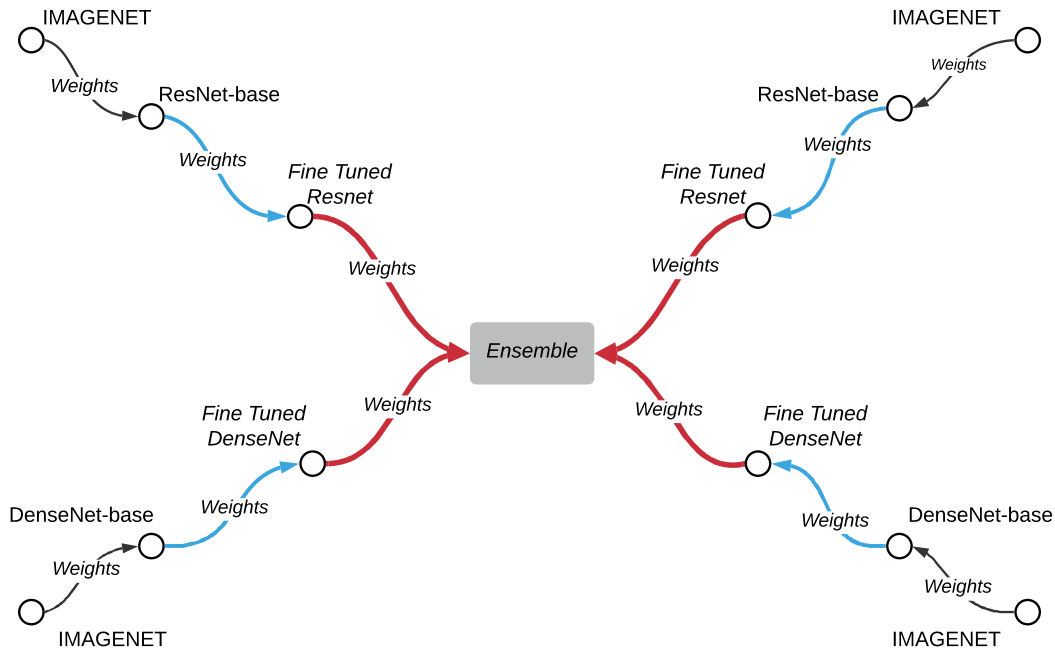


Fig. 1. Proposed PlexNet Architecture.

matching is based on error correcting code that may produce incorrect results. Further, if the key is leaked, the original template can easily be acquired.

Fuzzy vault and fuzzy commitment are cryptographic based protection techniques [20], [12]-[13]. A fuzzy vault scheme where a key k is locked using an unordered set A is proposed by Juels and Sudan [20]. During the enrollment phase, a polynomial p comprising of the biometric features is used to encode key k resulting in $p(k)$. The unordered set A is projected on p and some false points are added to secure the location of p . During authentication, if the query template overlaps the genuine location of p , key is reconstructed. Further, the idea was extended by Uludag *et al.*, securing fingerprints [21]. Nandakumar *et al.*, proposed the idea of using password for further strengthening fuzzy vault [22]. Though a robust technique in the past, this scheme fails to provide revocability and protection from matching across biometric systems.

Another key binding approach *i.e.*, fuzzy commitment is developed combining two approaches of cryptography and error correcting code. Here, during enrollment phase, a random key is encoded using error-correcting codes. The resultant is a codeword c . A hash value of c is also computed. The XOR operation is performed between generated codeword c and the biometric features, resulting in an encrypted template. This encrypted template known as helper data, is stored along with the hash value of c . During authentication, the features from

query template are decoded and XOR operation with helper data is performed to obtain codeword c' . If hash c matches the hash c' , the user is accepted as genuine. Further, this scheme is extended using helper data with face biometric [15]. The limitation of the fuzzy commitment scheme lies in the use of the template as helper data. If the biometric data is non-uniform, helper data may leak information on the secret key as well as about the template itself [14].

(b) *Key generation approach:* In this approach, the helper data is generated using biometric template. The key is generated from the helper data and query template. The basic principle of key generation technique is the quantization of helper data that produces the cryptographic keys [24], [26]. The biometric feature vector is divided into several intervals of feature elements. These intervals are selected from multiple biometric instances and encoded to store as helper data. During authentication, a similar approach is applied to the query template. A correct mapping between the intervals results in a key. The key generation algorithm with a quantization approach, can extract same keys from several instances of biometric data provides high stability. However, making it vulnerable to security attacks due to record multiplicity [19]. Moreover, if separate keys are generated for several instances of biometric data, the system suffers from increased entropy and reduced stability resulting in a high rejection rate.

The concept of secure sketch and fuzzy extractor

was introduced by Dodis *et al.* [25]. A secure sketch is obtained when a biometric template is encoded with a special algorithm. The query template and the secure sketch is required to decode the template, generating the near matching original template. Fuzzy extractors take input biometric identifiers as noisy data and generates a key as well as a helper data. During authentication, it requires both the helper data and the noisy data (query biometric) and generates the matching key.

2. *Template Transformation Approaches:* Here, a transformation function (f) is applied to the original biometric template (t) and only the transformed template [$f(t)$] is stored in the database. The transformation function can be derived from any random key or password-based approach. During matching, the same function (f) is applied to the query template (t') and is compared against the original transformed template [$f(t)$]. The template transformation can be further categorized into invertible *i.e.*, salting and non-invertible transformations [8], [49].

(a) *Salting technique:* In salting, the biometric features are transformed using an invertible transformation function. Since a key is used for transformation, it should be secured and only be available for authentication. Due to a fixed key used for transformation, system achieves low false acceptance rate (FAR). Connie *et al.*, used pseudo random keys for hashing palmprint templates [30]. A salting technique for face templates is proposed by Teoh and Ngo [31]. They also introduced a biophasor technique based on salting [28]. The transformed template can be inverted into the original if the transformation function is known. Thus, the salting technique suffers from the problem of invertibility making it susceptible to adversary attacks. Random projection is another example of salting technique, where the original biometric template is projected to another domain by a matrix with zero mean and unit variance. The problem with random projection is that the original matrix can be recovered if the projection matrix is known to the adversary. Zhe *et al.*, proposed a ranking based hashing that demonstrated two notions, one used random projection while the other used random permutation-based hashing schemes. Hence, generating invertible as well as non-invertible transformations [32].

(b) *Non-Invertible transformation:* On the contrary, a key is used as a transformation function in the non-invertible transformations making it a one-way transform. So, even if the transformation function is known, the inversion is computationally harder to achieve. A non-invertible transformation function for fingerprint is used by Ratha *et al.* [34]. In another work, they proposed Cartesian, polar and functional transforms that are non-invertible [11]. These functions transform the fingerprint minutiae *e.g.*, the Cartesian transformation divides the minutiae space into a grid and each cell is rearranged to a new position according to the key. The query template is transformed similarly and matched to the original transform for authentication. The use of robust hash function provides better cancelability

and security than salting, but the performance may decrease. One of the reasons for low performance of these biometric systems are that they do matching in transformed domain that results in high FRR. Further, if the transformation function is compromised, the system loses non-invertability. Thus, it is difficult to achieve a balance between discriminability and non-invertability.

3. *Hybrid Approaches:* Hybrid approaches were also introduced to take the advantages of both biometric cryptosystem and transformation. Feng *et al.*, proposed an approach, where the extracted template is randomly projected to a subspace generating a cancelable template [33]. The discriminability is increased during the transformation and finally fuzzy commitment is used to protect the generated cancelable template. A hybrid method based on revocable biotokens for face biometrics is proposed by Boulton [35]. Liu *et al.*, combined fuzzy vault technique with multi-space random projection to improve the security of palmprints [36]. The fingerprint and palmprint features are combined to build a hybrid system [37]. Further, neural network was introduced in combination with secure sketch [38]. This combination provides high tolerance against noisy data as well as revocability and non-invertibility. Sardar *et al.*, proposed a hybrid solution by combining the non-invertibility with a biocode encryption. This helped in enhancing security levels along with better accuracy [39]. Talreja *et al.* proposed a multibiometric template protection using binarization and hashing that uses a key for each individual during identification, hence the use had to have they key during the recognition process [40].

III. PROPOSED ENSEMBLE FRAMEWORK

The machine learning techniques simulate the learning process of the human brain and automatically create significant feature vectors. The major issue of these techniques is their limited generalization capability. For example, if the unseen patterns are to be tested, these methods usually results in unconvincing predictions. To improve generalizability, deep neural networks, particularly convolutional neural networks (CNN's) have evolved. The CNN's are proved to be very effective in several image processing and computer vision applications [44]. The capabilities of CNN's are yet to be explored for biometric template protection. The performance and generalizability of CNNs motivated us to utilize it for securing biometric templates. Two state-of-the-art CNN architectures *i.e.*, ResNet and DenseNet are used in order to form an ensemble.

A. CNN Architectures

The CNN is designed stacking three types of layers *i.e.*, convolution, pooling and a dense or fully connected layer. The convolution layer containing filters with numeric weights is responsible for feature extraction. The input image is convoluted with filters generating feature map. It contains the dominant features while preserving the relationship between the neighbouring pixels. More formally, let the input image be denoted as x , the filter be f and the rows and columns of the

convoluted matrix are denoted as a and b , respectively. Then, the convolution operation is defined as,

$$\text{Conv}[a, b] = (x * f)[a, b] = \sum_i \sum_j (f[i, j] x[a - i, b - j]) \quad (1)$$

Two CNN architectures *i.e.*, ResNet and DenseNet are chosen for making an ensemble [41], [42]. The selection of these architectures is based on their performances on VGGFace2 and MegaFace databases. In addition, these two architectures have several variants and proved to be complimentary to each other.

1) *ResNet Architecture*: The main idea behind ResNet is the skipping of one or more layers. The convolutional, relu, pooling and batch normalization layers are added in a redundant manner to form a residual network. The initial block consists of two 3×3 convolutional layers where each convolutional layers comprise 64 filters of size 3×3 . The rectified linear activation (ReLU) is the default activation followed by a batch normalization to maintain the mean activation as zero and the standard deviation closer to 1. The size of the second convolutional layer is reduced using max-pooling of size 2×2 . After each convolutional layer a dropout is applied to avoid over-fitting [50]. The strength of ResNet lies in its identity short-cut connection that flows the gradient by skipping one or more layers. It solves the vanishing gradient problem where the back-propagated gradient becomes infinitely small due to repeated multiplication. Thus, the identity short-cut connection provides the way to feed-forward the output of a ResNet block directly to the other block in the next layer.

More formally, let $\Theta_l(\cdot)$ be a non-linear composite transformation function of convolution, ReLU, pooling and batch normalization. The output of a ResNet block at layer l taken as Y_l , is [41],

$$Y_l = \Theta_l(Y_{l-1}) + Y_{l-1} \quad (2)$$

where Y_{l-1} , represents the input to a ResNet block at layer l . It means output of a ResNet block is depends on input from previous layer and it's non-linear transformation with function $\Theta_l(\cdot)$. Several such blocks are stacked that create a short-cut path for the gradient that optimizes the back-propagation. Due to stacking, ResNet may results in generating redundant layers.

2) *DenseNet Architecture*: Huang *et al.*, proposed another solution to vanishing gradient problem using DenseNet [42]. DenseNet comprises of several blocks that are densely connected. The features generated by one block is input to the next dense block. In larger CNN models, there may be feature loss before it reaches to the output layer. DenseNet overcomes this issue with the use of repeated blocks. Further, the DenseNet requires fewer parameters in comparison to the other models such as ResNets, [43].

DenseNet architecture consists of dense block and transition layers. The dense block is a collection of different type of layers connected to similar previous layers. The feature maps are generated using 3×3 convolutional layer within a block. The generated feature maps are scaled with a batch normalization layer. The feature maps generated by each layer are concatenated together. Thus, the output of a layer l in a DenseNet unit can be represented as:

$$Y_l = \Theta_l([Y_0, Y_1, \dots, Y_{l-1}]) \quad (3)$$

Here, the transition between two dense layers is achieved through a transition layer that controls the number of connections. The ResNet and DenseNet architectures are found to be complementary to each other due to the fact that both provide solution to vanishing gradient. The ResNet handles it with large number of parameters that is compensated by feature reuse in DensNet.

B. Transfer Learning

In order to overcome the issue of under-fitting in smaller databases, the concept of transfer learning is introduced [51]. The weights of a pre-trained model are utilized while training a new model on a different database. Thus, the learning starts from an elevated point that avoids data insufficiency. Moreover, the identical distribution of training and test data is not required while using transfer learning.

Formally, let, ζ_1 be the learning curve of a model pre-trained on database χ_1 . The goal is to improve the learning curve, ζ_2 for a new model with a new database χ_2 . The learnt behaviour of the model pre-trained on, χ_1 is transferred to the predictive function, ϕ . The function ϕ in turn, improves the learning curve, ζ_2 on database χ_2 . Several state-of-the-art image classification methods are based on transfer learning [53]. Different model are pre-trained on ImageNet that contains millions of images of over 1000 categories [18]. The upper layers of these pre-trained models can be fine tuned to match the current model working on the new database.

C. Ensemble Learning

The pre-trained models perform better than a model that starts learning from the scratch. However, sometimes a single pre-trained model may not be enough to establish a robust learning for a given database due to class imbalance or concept drift [52]. To overcome this issue, ensemble of learning algorithms is used for many classification problems [54]. It performs learning by combining different learning models to a single predictive model.

An ensemble network can be built using different classifier architectures, initial weights and training databases. Although, the use of different classifier architecture may be a good choice, it requires compatible architectures. For example, ResNet-152 architecture may accept the initial information from a ResNet-101 or ResNet-50, but can not accept the initial weights from other architectures such as DenseNet or VGG [41], [42].

The proposed ensemble framework *i.e.*, PlexNet is prepared using two pre-trained architectures *i.e.*, ResNet and DenseNet. These architectures are chosen over other architectures such as VGG, MobileNet, ResNetV2 and InceptionResNetV2, due to the following reasons,

- The ResNet and DensNet architectures have several variants that are compatible to each other. Thus, an ensemble with different classifier architectures is achieved.
- Secondly, these architectures have achieved better accuracies than other pre-trained architectures as shown in Table I.

The initial training starts with pre-trained models on ImageNet database [45]. These initial models are presented as the endpoint of the framework as shown in Fig. 1. The weights obtain after training the initial models on the query database are saved. These weights are then fed to the next level of compatible architectures present at the axons of the PlexNet architecture. Thus, the axon models are fine-tuned and achieve higher accuracy with faster convergence than their predecessors. Finally, these fine-tuned models are combined in an integrated stacking model to prepare a meta-learner *i.e.*, Now, feature level fusion is performed that fuses features from each fine-tuned model using concatenation. The result of fusion is a feature vector representing weights from each class. Further, the features are passed to a fully connected neural network of ensemble model that map the features to their corresponding classes based on expected likelihood.

The robust learning of PlexNet that ensures lower false prediction is collected and saved in a smart box *i.e.*, known to be an API as shown in Fig. 2. The query image does not expose features but maps to the appropriate class while processed with API. The model resembles the learning of a human brain to recognise a person using facial images. Learning with faces, again and again, created patterns in the brain that are to be known. If a similar pattern occurs then, there is no need to seek any reference but the brain matches the description with the learning did in the past and identify the person. Similarly, the API predicts the appropriate class of a query image referencing it's learning. Therefore, our model of template protection need not store templates in database for making predictions. Thus, the requirement of nurturing a template database is eliminated, hence making it impregnable against fraudulent attacks.

D. Data Augmentation

To reduce the over-fitting due to class imbalance or rather unavailability of sufficient data, augmentation of images is performed for accommodating the intraclass variations [46]. Various transformation techniques are applied on the available set of images, such as geometric (*e.g.*, reflection, scaling, rotation, shear, translation) and photometric (*e.g.*, brightness, noise reduction, hue adjustments, edge enhancements).

In order to maintain the heterogeneity of generated images and space-time complexities to be minimum, online data augmentation is introduced. We use the image 'Datagenerator' function available in Keras that apply different transformations in each epoch [47]. For example, the images are flipped in one epoch, whereas zooming is applied in another epoch. As the number of epochs increases, so the number of transformation on the random images per class. Thus, each model learns with a different set of images per class at the axon level.

IV. EXPERIMENTAL RESULTS

A. Databases

Two publicly available databases *i.e.*, VGGFace2 and MegaFace are used for training and testing all the architectures. The VGGFace2 database comprises of more than 9000 subjects spanning over different accents, ages and ethnicities [55]. There are more than 350 facial images per subject that are further subjected to augmentation. The size of the images varies from 50 pixels to more than 300 pixels, which are

TABLE I. TESTING ACCURACIES OF BASE MODELS USING FACE DATABASES

Architecture	Accuracy (%)			
	VGGFace2		MegaFace	
	80:20	90:10	80:20	90:10
ResNet	93.14	90.93	96.46	94.65
DenseNet	92.52	89.12	97.21	95.74
MobileNet	88.71	85.41	95.39	94.18
VGG	89.25	87.34	96.29	95.96
ResNetV2	90.69	88.32	95.76	92.77
InceptionResNetV2	89.26	86.91	95.93	93.34

averaged to 225 pixels for the experimentation limitations. The facial images of this database are acquired in an unconstrained environment. It means the images have pose and style variations, front or side views and may contain torso part, hence putting more challenges to a face recognition system [56].

The MegaFace database contains millions of facial images belonging to hundred of classes [57]. The images are compressed using JPEG formats. The three channel colored images are available with a dimension range from 250×226 to 300×312 . These are also changed from their original size to 225×225 . The classes have an unbalanced distribution of images ranging from 500 to 700. In MegaFace database, images are taken in a constrained environment *i.e.*, the faces are cropped and only the front view is included.

B. Validation Metrics

Let, $Pred = (Pred_1, Pred_2, Pred_3 \dots Pred_n)$ be the predicted classes by the classification algorithm, and $Act = (Act_1, Act_2, Act_3 \dots Act_n)$ be the actual n classes. The proposed method is evaluated using following validation metrics,

- **True Positives (TP):** is the representation of a correct prediction of a positive class by the model *i.e.*, ($Pred_i = Act_i$), where i represents class labels.
- **True Negatives (TN):** similar to that of true positives, true negatives are the correct identification of the negative class by the model *i.e.*, ($Pred_i \in (1 - Act)$).
- **False Positives (FP):** If the model predicts the negative class as the positive class it is termed as false positives *i.e.*, ($Pred_i = Act_j$), where i, j represents class labels from predictions and actual class sets, respectively and $i \neq j$.
- **False Negatives (FN):** If there is an incorrect prediction of a negative class it's a false negative *i.e.*, ($(1 - Pred_i) \in (Act)$).
- **Confusion Matrix:** From the above predicted and actual labels, confusion matrix C can be derived for the classes where $C_{i,i}$, determines the correct predictions for the i^{th} class whereas $C_{i,j}$ determines the i^{th} classes that were misclassified as j^{th} class.
- **Precision & Recall:** With the calculation of number of TPs, FPs and FNs, the precision and recall are calculated as,

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

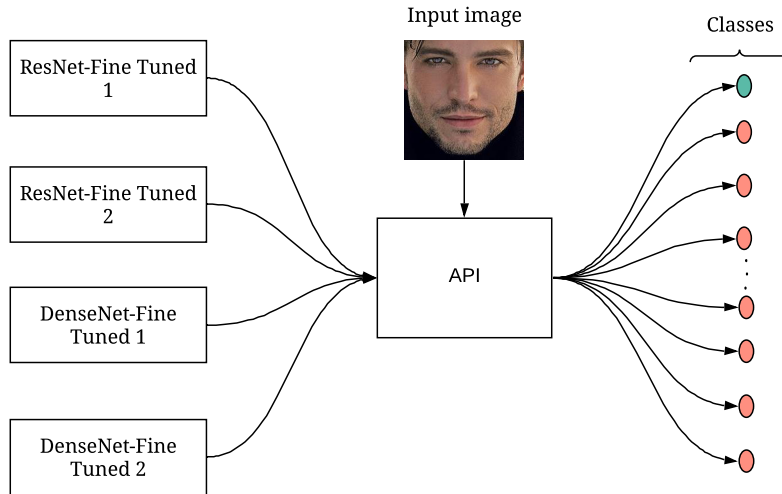


Fig. 2. API Generation using Score Fusion and Classification.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

- *F1-Score*: is the harmonic mean that represents a single measure for both precision and recall.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

C. Results

The selection of pre-trained models in PlexNet is based on the accuracy achieved on VGGFace2 and MegaFace databases, as shown in Table I. The accuracies of different architectures are evaluated at two different training and testing database ratio *i.e.*, 80:20 and 90:10. The accuracies at the 80:20 split ratio are found better than the accuracy at split ratio of 90:10 on both databases. So, all the experiments are performed on the training and testing database ratio of 80:20. On VGGFace2 database, the accuracy of ResNet base model is found to be the highest *i.e.*, 93.14% at the training and testing database ratio of 80:20. Similarly, the DenseNet achieves the best accuracy of 92.52% at the base model. The accuracies of all other tested architectures are found lower in comparison to these models. Similarly, on MegaFace database these two architectures achieve better accuracies. For example, the ResNet and DenseNet architectures achieve the accuracies of 96.46% and 97.21%, respectively. Beside the accuracies, ResNet and DenseNet complement each other, hence chosen as base models.

The PlexNet is prepared through the integrated stacking model using pre-trained architectures. The information gathered using all the pre-trained weights helps the ensemble to attain the best accuracies as shown in Fig. 3. The base models chosen for both the databases achieve lower accuracies. On VGGFace2 database, the ResNet base model achieves the accuracies of 88.54%, 93.45% and 93.14% at epochs 1, 5 and 10, respectively as shown in Fig. 3a. Similarly, the DenseNet base model at epochs 1, 5 and 10, reports the

accuracies 73.24%, 90.12% and 92.52%, respectively. The fine-tuned model perform better using the learning weights of base models. For example, the accuracies of ResNet and DenseNet fine-tuned models are found to be 94.91% and 93.95%, respectively at 10 epochs. Finally, the fusion of fine-tuned models results in PlexNet that reports the accuracy of 96.48% on the VGGFace2 database.

The proposed method performs better on MegaFace database. The testing accuracies on MegaFace database is shown in Fig. 3b. Here, the testing accuracies of ResNet base model are 82.66%, 94.56% and 96.44% at 1, 5 and 10 epochs, respectively. The DenseNet base model performs better than ResNet base model. It reports the accuracies of 81.65%, 95.12% and 97.21% at 1, 5 and 10 epochs, respectively. Since, the MegaFace database contains facial images acquired in a constrained environment, it may results in lower intraclass variations, hence better accuracies are achieved using all the models. Here, the ResNet and DenseNet fine-tuned models reported the highest accuracies of 97.57% and 98.16% at 10 epochs, respectively. The overall accuracy of proposed PlexNet on MegaFace database outperforms all other existing methods and found to be 99.61%.

The values for precision, recall and F1-score on MegaFace database found to be better than VGGFace2 database and reported to 0.9948, 0.9946 and 0.9947, respectively. The higher values reported for these metrics on both databases further show the robustness of PlexNet. The values for these metrics demonstrate that VGGFace2 is a challenging database, as it contains images in unconstrained environment with large variations in pose, illumination and ethnicity.

The effectiveness of a biometric system is evaluated in terms of true acceptance rate (TAR). In order to achieve better visualization of the accuracy of the PlexNet, a receiver operating characteristic (ROC) curve is drawn as shown in Fig. 4. The ROC curve plots false acceptance rate (FAR) on x-axis and TAR on y-axis. Thus, a relationship is established between FAR and TAR that can be utilized while deciding a

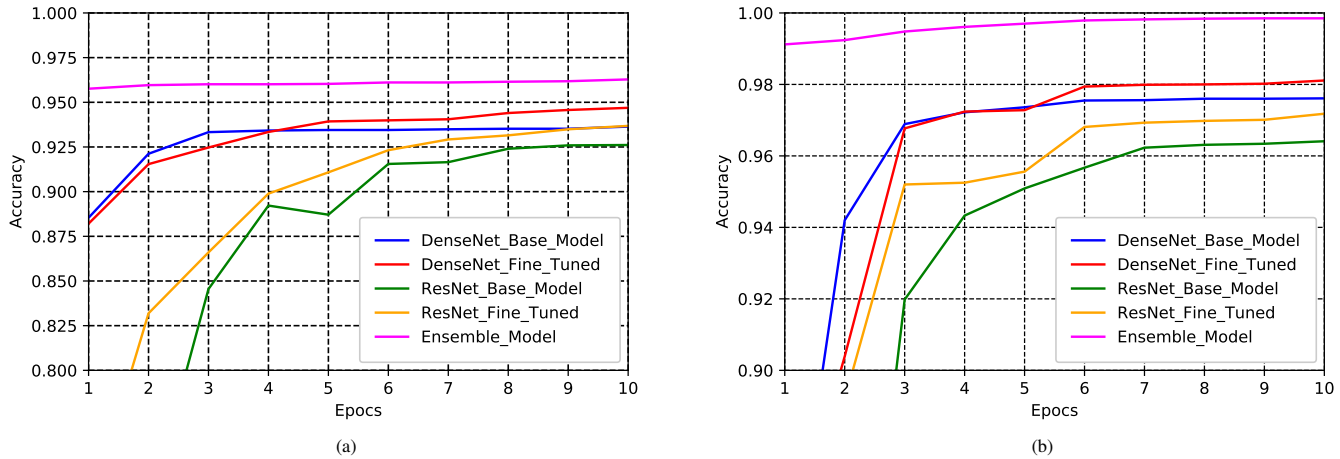


Fig. 3. Visualization of Testing Accuracies using a) VGGFace2 and b) MegaFace Databases.

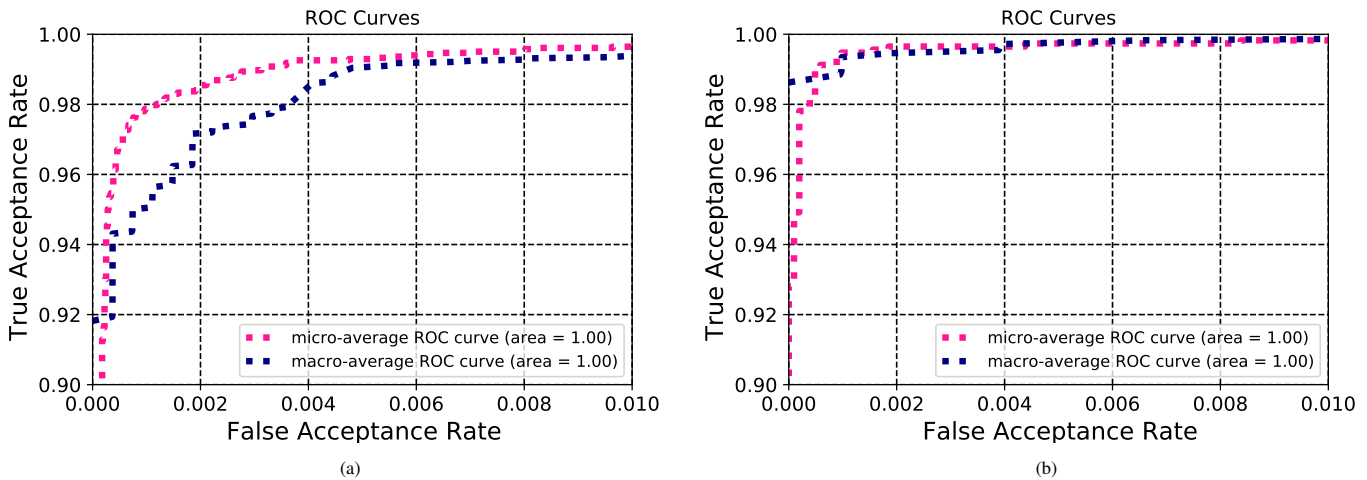


Fig. 4. Receiver Operating Characteristic Curves for a) VGGFace2 and b) MegaFace Databases.

threshold for FAR to attain optimal TAR. The PlexNet achieves 98% and 99% TAR at 0.2% and 1% FAR, respectively on the VGGFace2 database. It shows the effectiveness of PlexNet even on databases acquired in an unconstrained environment. The PlexNet performs better on the MegaFace database and reports a higher TAR of 99.6% just at 0.2% FAR. The TAR reaches to 100% at the FAR value of only 0.8%.

D. Comparative Analysis

The performance of proposed template protection method is compared with existing methods on various biometric databases as shown in Table II. The template protection methods using fingerprint biometrics reported the GAR of 97% and 94% at 0.1% and 0% FAR, respectively [62], [58]. The discriminability of iris features proved to be better as the method of Talreja *et al.* reported 99.1% GAR at 0% FAR [40]. However, the iris patterns are taken obtrusively and are inconvenient for users. Therefore, the recognition

TABLE II. PERFORMANCE COMPARISON WITH EXISTING METHODS OF TEMPLATE PROTECTION

Method	Dataset	GAR%@FAR%
Nagar et al. [62]	FVC 2002	97.0@0.1
Kumar et al. [63]	IITD iris	91.0@0.0
Talreja et al. [40]	WVU iris	99.1@0.0
Hybrid Approach [59]	Multi-PIE	90.61@1.0
BDA [60]	Multi-PIE	96.38@1.0
DeepCNN [61]	Multi-PIE	96.53@0.0
Wang & Hu [58]	FVC 2002	94.0@0.0
Our Method	VGGFace2	92.0@0.0
Our Method	MegaFace	98.5@0.0

performance of iris biometrics may deteriorate for real-time databases as in case of Kumar *et al.*, who reported 91% GAR at 0% FAR [63]. Most of the template protection methods tested on face biometrics use Multi-PIE database that is taken

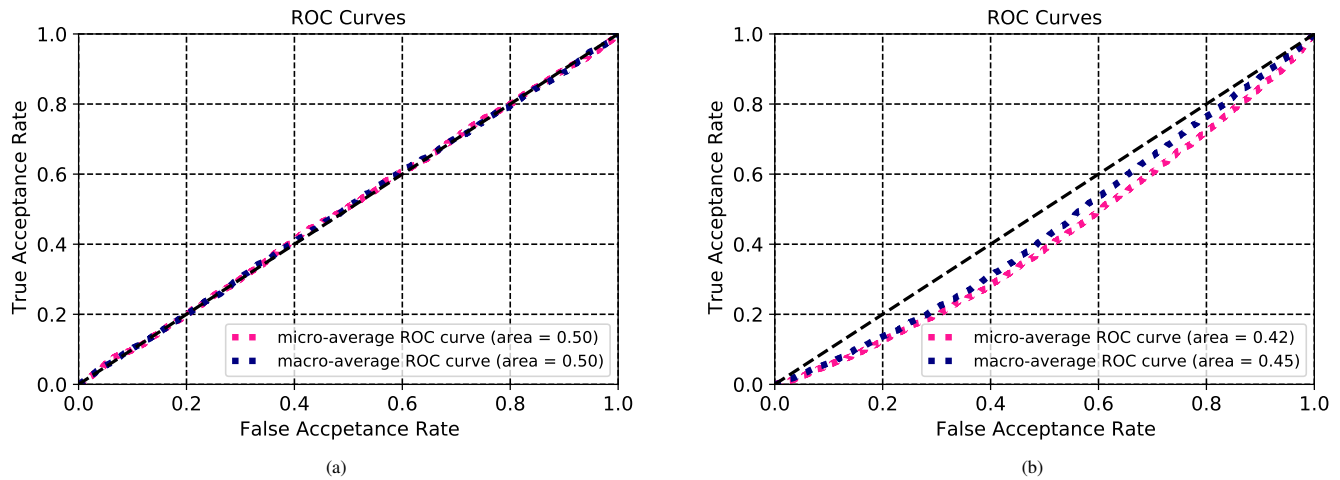


Fig. 5. ROC Curve for Birthday Attack on a) VGG and b) MegaFace Databases.

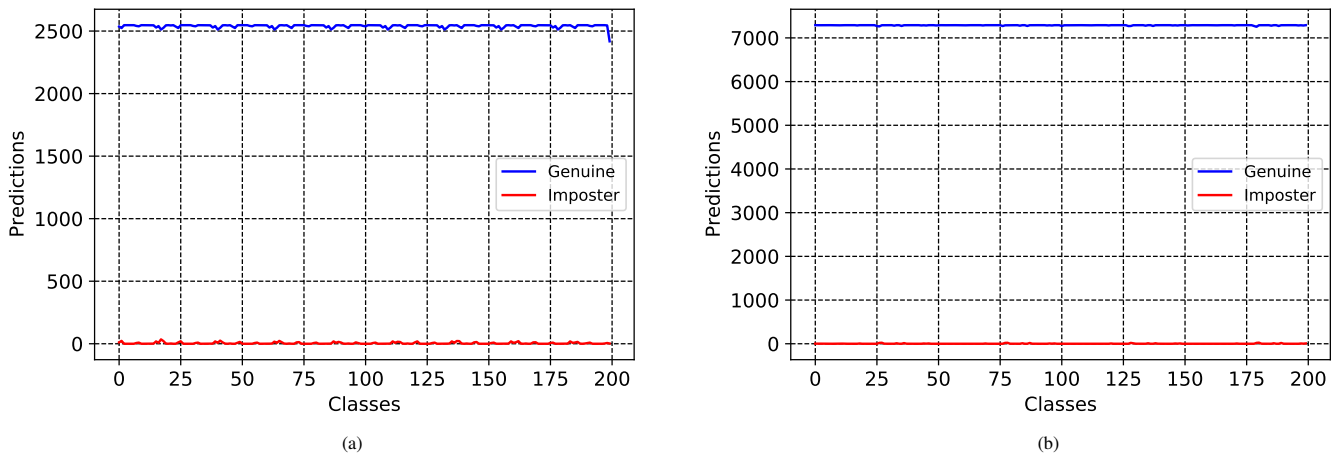


Fig. 6. Genuine Imposter Distribution for Birthday Attack on a) VGG and b) MegaFace Databases.

in constrained environment. The performance of proposed method is evaluated on both constrained and unconstrained databases. The proposed method reports 92% GAR at 0% FAR on VGGFace2 database that is taken in unconstrained environment. The performance of proposed method improves on MegaFace database. In comparison to other constrained environment databases, our method reports highest GAR of 98.5% at 0% FAR. Thus, the proposed method outperforms other existing methods on both constrained and unconstrained databases.

E. Security Analysis

The security paradigm of the proposed ensemble of biometric template security is analyzed to avoid possible fraudulent attacks on the biometric system. We can consider following parameters to evaluate the robustness of the proposed template protection method.

- **User Impersonation:** The attacker may try to impersonate as a genuine user. The experimental results show that higher TAR is achieved at marginal FAR on both databases. For example, on the VGG2Face database 99% TAR is achieved at lower FAR of 1%. It can be further reduced without much loss in TAR. The false acceptance is further minimized while using the MegaFace database, as 100% TAR is achieved at 0.8% FAR. The proposed biometric system is to be set as to not allow any false acceptance *i.e.*, FAR of 0%, while maintaining the TAR of 93% and 97% on VGGFace2 and MegaFace database, respectively. Hence, there is less chance of template impersonation while using the unconstrained database. Finally, the false biometric attempts made by an intruder to breach the security of the system can be negligible.
- **Denial to Attackers:** Generally, in a biometric system the features of a query image are matched with the

templates stored in the database. An attacker may gain access to the templates of the gallery in such cases. In the proposed framework of PlexNet the gallery database is removed. The learning of PlexNet is enough to make a correct prediction without maintaining the template database. Therefore, there is hardly any possibility of attacks on the template database. The proposed method predicts the class label for a query image using a smart box *i.e.*, API. The API may be compromised by an attacker. It consists of weights and biases from different pre-trained models having millions of parameters per model. The access of image information from these parameters is almost impossible. Hence, the proposed ensemble ensures the security of biometric templates that are completely irrevocable.

To further strengthen our claims, we performed birthday attack to check the vulnerability of the prepared model against cross-referencing of the database. To achieve this, both the databases were cross referenced against each other. For example, the API prepared using VGG face dataset was fed the MegaFaceDataset and vice-versa. The result of this attack is shown in Fig. 5 and Fig. 6. The ROC curve is considered most accurate if it is close to the top left of the corner and then moves horizontally. Likewise, if the plot is closer to the diagonal, as shown in Fig. 5a and 5b, it shows that the probability of false identification is very low. The birthday attack is further analyzed using positive and negative predictions. In both the cases, the positive predictions tend to be near 0 or exact 0 while negative predictions were closer to 1, as shown in Fig. 6a and 6b. This further strengthens the claim that the proposed model has negligible false acceptance during an adversary attack.

V. CONCLUSION AND FUTURE SCOPE

The deep neural networks have achieved tremendous performance in computer vision, pattern recognition and image analysis. These networks learn intrinsic patterns of data automatically without being programmed. The advantages of using deep neural networks such as CNNs are yet to be explored for the application of biometric template protection. The application of CNNs in biometric recognition has motivated us to utilize the potential for biometric template protection. It has shown that a single learning model may not perform better on complex databases due to class imbalance or data insufficiency. The class imbalance and data insufficiency may result in the problem of over-fitting and under-fitting, respectively. To overcome these issues, an ensemble has proved to be efficient in several state-of-the-art methods of biometric authentication. This paper has presented a detailed study on biometric template protection. It has been achieved through exclusion of templates from the database by exploiting the learning of deep neural networks. The exclusion of the template database requires a high performing network architecture, so an ensemble has designed.

The ensemble called 'PlexNet' is designed using deep neural networks. Two state-of-the-art CNN architectures *i.e.*, ResNet and DenseNet are chosen to form the PlexNet. The

selection of these architectures is based on experimentation on a database of millions of images and their complementarity. The selected base models are trained separately using their pre-trained weights from ImageNet. The performance of PlexNet for biometric template protection has evaluated on two publicly available databases *i.e.*, VGGFace2 and MegaFace. It has achieved outstanding recognition results that outperform other state-of-the-art template protection mechanism using face biometrics [56].

The proposed PlexNet model laid the foundation for biometric template protection. An API has prepared that learnt base and fine-tuned models together. The API is acted as a smart box for the query image that has predicted the correct class without exposing the biometric features during training. The proposed template protection framework did not store templates in the gallery for any possible predictions by intruders. Therefore, proposed framework has made possible the exclusion of templates from gallery and performed predictions based on learning that is irrevocable.

REFERENCES

- [1] A. Singh, R. Srivastva, Y. N. Singh, "Prevention of Payment Card Frauds using Biometrics", International Journal of Recent Technology and Engineering (IJRTE), 8 (3S), pp. 516-525, 2019.
- [2] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition", IEEE Transactions on Circuits and Systems for Video Technology, 14 (1), pp. 4-20, 2004.
- [3] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security", IEEE Transactions on Information Forensics and Security, 1 (2), pp. 125-143, 2006.
- [4] Y. N. Singh and S. K. Singh, "A Taxonomy of Biometric System Vulnerabilities and Defenses", International Journal of Biometrics, 5 (2), pp. 137-159, 2013.
- [5] Grottko M., Matias R. and Trivedi K.S., "The fundamentals of software aging", IEEE International Symposium on Software Reliability Engineering, pp.1-6, 2008.
- [6] A. K. Jain, R. Bolle, and S. Pankanti, Eds., "Biometrics: Personal Identification in Networked Society", Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
- [7] R. Srivastva, A. Singh, Y. N. Singh, "PlexNet: A fast and robust ECG biometric system for human recognition", Information Sciences, 558, pp. 208-228, 2021.
- [8] U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, "Biometric cryptosystems: issues and challenges", Proc. of the IEEE, 92 (6), pp. 948-960, 2004.
- [9] A. Cavoukian and A. Stoianov, "Biometric encryption: a positive-sum technology that achieves strong authentication, security and privacy", Tech. Rep., Office of the Information and Privacy Commissioner of Ontario, Toronto, Ontario, Canada, March 2007.
- [10] A. Vetro and N. Memon, "Biometric system security", in Proc. of the 2nd International Conference on Biometrics, Seoul, South Korea, August 2007.
- [11] N. K. Ratha, S. Chikkerur, J. H. Connell, and R. M. Bolle, "Generating cancelable fingerprint templates", IEEE Transactions on pattern analysis and machine intelligence, 29 (4), pp. 561-572, 2007.
- [12] A. Juels and M. Wattenberg, "A fuzzy commitment scheme", Proceedings of the 6th ACM conference on Computer and communications security, pp. 28-36. ACM, 1999.
- [13] H. Lu, K. Martin, F. Bui, K. N. Plataniotis and D. Hatzinakos, "Face recognition with biometric encryption for privacy-enhancing self-exclusion", 2009 16th International Conference on Digital Signal Processing, Santorini-Hellas, pp. 1-8, 2009.
- [14] P. Tuyls and J. Goseling, "Capacity and examples of template-protecting biometric authentication systems", In: Maltoni D., Jain A.K. (eds) Biometric Authentication. BioAW 2004. Lecture Notes in Computer Science, vol 3087. Springer, Berlin, Heidelberg, pp. 158-170, 2004.

- [15] M. VanderVeen, T. Kevenaar, G. J. Schrijen, T. H. Akkermans, F. Zuo, "Face biometrics with renewable templates", in Proc. of SPIE: Security, Steganography, and Watermarking of Multimedia Contents, vol. 6072, 2006.
- [16] K. Nandakumar, A. Jain, "Biometric template protection: bridging the performance gap between theory and practice", IEEE Signal Processing Magazine, 32 (5), 88-100, 2015.
- [17] Y. Sutcu, Q. Li, N. Memon, "Protecting biometric templates with sketch: theory and practice", IEEE Transactions on Information Forensics and Security, 2 (3), pp. 503-512, 2007.
- [18] Alex Krizhevsky, Ilya Sutskever, G. E. Hinton, "Imagenet Classification with Deep convolutional neural network", Communications of the ACM, 60 (6), May 2017.
- [19] C. Rathgeb, A. Uhl and P. Wild, "Iris-biometrics: from segmentation to template security", S. Jajodia (ed.), Advances in Information Security, Springer, 2013.
- [20] A. Juels and M. Sudan, "A fuzzy vault scheme. Designs, Codes and Cryptography", 38 (2), pp. 237-257, 2006.
- [21] U. Uludag, S. Pankanti, A. K. Jain, "Fuzzy vault for fingerprints", in Proc. of 5th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA 2005, Hilton Rye Town, NY, USA, 20-22, July 2005.
- [22] K. Nandakumar, A. Jain, S. Pankanti, "Fingerprint-based fuzzy vault: implementation and performance", IEEE Transactions on Information Forensics and Security 2 (4), pp. 744-757, 2007.
- [23] Dang, T. K., V. Q. P. Huynh, and Q. H. Truong, "A Hybrid Template Protection Approach using Secure Sketch and ANN for Strong Biometric Key Generation with Revocability Guarantee". The International Arab Journal of Information Technology (IAJIT), 15 (2), pp. 331-340, 2018.
- [24] Y.-J. Chang, W. Zhang, and T. Chen, "Biometrics-based cryptographic key generation", in Proc. of the IEEE International Conference on Multimedia and Expo (ICME '04), vol. 3, pp. 2203-2206, Taipei, Taiwan, June 2004.
- [25] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy extractors: how to generate strong keys from biometrics and other noisy data", Tech. Rep. 235, Cryptology ePrint Archive, February 2006.
- [26] C. Vielhauer, R. Steinmetz, and A. Mayerhofer, "Biometric hash based on statistical features of online signatures", in Proc. of the International Conference on Pattern Recognition, vol. 1, pp. 123-126, Quebec, QC, Canada, August 2002.
- [27] D. C. L. Ngo, A. B. J. Teoh, J. Hu, "Biometric Security", Cambridge Scholars Publishing, Newcastle upon Tyne, 2015.
- [28] A. B. J. Teoh, A. Goh, and D. C. L. Ngo, "Random multispace quantization as an analytic mechanism for BioHashing of biometric and random identity inputs", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (12), pp. 1892-1901, 2006.
- [29] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, "Handbook of Fingerprint Recognition", Springer, Berlin, Germany, 2003.
- [30] T. Connie, A. Teoh, M. Goh, D. Ngo, "Palmhashing: a novel approach for cancelable biometrics", Information Processing Letters, 93 (1), 1-5, 2005.
- [31] A. Teoh, D. Ngo, "Biophasor: token supplemented cancellable biometrics", in 9th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 1-5, 2006.
- [32] Z. Jin, J. Y. Hwang, Y. Lai, S. Kim and A. B. J. Teoh, "Ranking-Based Locality Sensitive Hashing-Enabled Cancelable Biometrics: Index-of-Max Hashing," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 2, pp. 393-407, Feb. 2018.
- [33] F. Y. Cheng, Y. P. Chi, and A. K. Jain, "A hybrid approach for generating secure and discriminating face template", IEEE Transactions on Information Forensics and Security, 5 (1), pp. 103-117, 2010.
- [34] N. K. Ratha, J. H. Connell, R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems", IBM Systems Journal, 40 (3), pp. 614-634, 2001.
- [35] T. Boulton, "Robust distance measures for face-recognition supporting revocable biometric tokens", in 7th International Conference on Automatic Face and Gesture Recognition (FGR), pp. 560-566, 2006.
- [36] H. Liu, D. Sun, K. Xiong, Z. Qiu, "A hybrid approach to protect palmprint templates". Scientific World Journal 2014, pp. 686-754, 2014.
- [37] Y. Chin, T. Ong, A. Teoh, K. Goh, "Integrated biometrics template protection technique based on fingerprint and palmprint feature-level fusion". Information Fusion 18, pp. 161-174, 2014.
- [38] Dang, T. K., V. Q. P. Huynh, and Q. H. Truong, "A Hybrid Template Protection Approach using Secure Sketch and ANN for Strong Biometric Key Generation with Revocability Guarantee", The International Arab Journal of Information Technology (IAJIT), 15 (2), pp. 331-340, 2018.
- [39] A. Sardar, S. Umer, C. Pero and M. Nappi, "A Novel Cancelable FaceHashing Technique Based on Non-Invertible Transformation With Encryption and Decryption Template," in IEEE Access, vol. 8, pp. 105263-105277, 2020.
- [40] V. Talreja, M. C. Valenti and N. M. Nasrabadi, "Deep Hashing for Secure Multimodal Biometrics," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1306-1321, 2021.
- [41] 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), Las Vegas, NV, 2016.
- [42] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp. 2261-2269, 2017.
- [43] Y. M. Glaser, "Densely Connected Convolutional Neural Networks for Natural Language Processing", Honors Theses 36, University of North Georgia, 2018.
- [44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, pp. 1-14, 2015.
- [45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 248-255, 2009.
- [46] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning", ArXiv e-prints, 2017.
- [47] Keras Applications, "<https://keras.io/applications/>", accessed 15th July 2020.
- [48] Cavoukian A., Stoianov A, "Biometric Encryption", In: van Tilborg H.C.A., Jajodia S. (eds) Encyclopedia of Cryptography and Security". Springer, Boston, MA, 2011.
- [49] A. Nagar and A. K. Jain, "On the security of non-invertible fingerprint template transforms", 2009 First IEEE International Workshop on Information Forensics and Security (WIFS), London, pp. 81-85, 2009.
- [50] Caruana, Rich & Lawrence, Steve & Giles, C, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", Advances in Neural Information Processing Systems. 13. pp. 402-408, 2000.
- [51] Lin, Jianzhe & Wang, Qi & Ward, Rabab & Wang, Z.. (2018). "DT-LET: Deep Transfer Learning by Exploring where to Transfer". Neurocomputing, Vol 390, pp. 99-107, 2020.
- [52] Ghaemi, R., Sulaiman, M. N., Ibrahim, H., & Mustapha, N., "A survey: Clustering ensembles techniques", World Academy of Science, Engineering and Technology, 50, pp. 636-645, 2009.
- [53] Tan, Chuanqi & Sun, Fuchun & Kong, Tao & Zhang, Wenchang & Yang, Chao & Liu, Chunfang, "Survey on Deep Transfer Learning", 27th International Conference on Artificial Neural Networks, Rhodes, Proceedings, Part III., 2018.
- [54] Yuan X, Xie L, Abouelenien M., "A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data", Pattern Recognition, 77, pp. 160-172, 2018.
- [55] Cao, Q. and Shen, L. and Xie, W. and Parkhi, O. M. and Zisserman, I. A., "VGGFace2: A dataset for recognising faces across pose and age", International Conference on Automatic Face and Gesture Recognition, 2018.
- [56] R. Shyam and Y. N. Singh, "Recognizing Individuals from Unconstrained Facial Images", Intelligent Systems Technologies and Applications, Advances in Intelligent Systems and Computing, Springer, vol. 384, pp 383-392, 2015.

- [57] Kemelmacher-Shlizerman, Ira and Seitz, Steven M and Miller, Daniel and Brossard, Evan, "The MegaFace benchmark: 1 million faces for recognition at scale", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873-4882, 2016.
- [58] Song Wang, Jiankun Hu, "A blind system identification approach to cancelable fingerprint templates", Pattern Recognition, Vol. 54, pp. 14-22, 2016.
- [59] Feng, Y. C., Yuen, P. C. and Jain A. K., "A hybrid approach for generating secure and discriminating face template", IEEE transactions on information forensics and security, 5(1), pp.103-117, 2010.
- [60] Feng Y. C. and Yuen, P. C. "Binary discriminant analysis for generating binary face template", IEEE Transactions on Information Forensics and Security, 7(2): pp. 613-624, 2012.
- [61] Jindal, A. K., Chalamala, S., and Jami, S. K., "Face Template Protection Using Deep Convolutional Neural Network", IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, pp. 575-5758, 2018.
- [62] A. Nagar, K. Nandakumar, A.K. Jain, "A hybrid biometric cryptosystem for securing fingerprint minutiae templates", Pattern Recogn. Lett. 31(8), 733-741 (2010)
- [63] A. Kumar, A. Kumar, "A cell-array based multibiometric cryptosystem". IEEE Access 4, pp. 15-25 2016

A Multi-layer Machine Learning-based Intrusion Detection System for Wireless Sensor Networks

Nada M. Alruhaily¹, Dina M. Ibrahim²

Department of Information Technology, College of Computer, Qassim University, Buraydah, Saudi Arabia^{1,2}
Department of Computers and Control Engineering, Faculty of Engineering, Tanta University, Egypt¹

Abstract—With the increase relay on the internet, and the shift of most business to provide remote services, the burdens of protecting the network and detecting any attack quickly become more significant, as the attack surface and Cyberattack increases in return. Most current Wireless Sensor Networks (WSNs) intrusion detection models that use machine learning methods to identify non-previously seen attacks utilize one layer of detection, meaning that a costly algorithm should be run before detecting any suspicious activity. In this paper, we propose a multi-layer intrusion detection framework for WSN; in which we adopt a defense-in-depth security strategy, where two layers of detection are deployed. The first layer is located on the network edge sensors are distributed; it uses a Naive Bayes classifier for real-time decision making of the inspected packets. The second layer is located on the cloud and utilizes a Random Forest multi-class classifier for an in-depth analysis of the inspected packets. The results demonstrate that our proposed multi-layer detection model gives a relatively high performance of the TPR, TNR, FPR, and FNR, additionally achieving a high Precision rate with values of, 100%, 90.4%, 99.5%, 97%, 99.9% for the Normal, Flooding, Scheduling, Grayhole, and Blackhole attacks, respectively.

Keywords—Intrusion detection; wireless sensor networks; machine learning; defence in depth strategy

I. INTRODUCTION

With the emergence of wireless devices, especially in the Wireless Sensor Networks (WSN), and due to the rapid spread of the Internet of Things technology, this has led to a dramatic increase of the attack surface resulting in the network being exposed to various types of attacks [1]. For this reason, intrusion detection methods with highly stability, efficiency, and adaptability are in urgent need to protect such networks. At present, the traditional wireless network intrusion detection methods suffer from some limitations like: low detection accuracy, low precision rate, and high false positive rate [2]. Therefore, there is a growing need to propose a more accurate and efficient intrusion detection framework to enhance the intrusion detection qualification in the wireless sensor network environment.

Nowadays, the application of artificial intelligence methods to intrusion detection systems has become one of the most important research fields carried out by researchers, especially using machine learning algorithms. Additionally, some researches are applying other methods including neural networks [3], [4], [5], genetic algorithms [6], [7], and deep learning techniques [8], [9], [10].

Most of the current frameworks proposed for detecting intrusions in Wireless sensor Networks deal with the network

as a whole; thus, they tend to propose one layer of detection, while WSNs consist of a considerable number of sensors distributed in a large area, as the works done by [1], [2], [3]. Therefore, our target in this paper is to divide the task of detecting the network intrusions between two detection layers. Where in the first layer, a simple classifier that has a very low computational cost (i.e. Naive bayes) is used to filter the malicious traffic and pass it to the second layer in which more extensive processing is carried out by utilizing a multi-class Random Forest classifier [11]. In the last few years, many approaches have been proposed to design intrusion detection systems for wireless sensor networks. authors in [12] introduced an evolutionary mechanism to extract intrusion detection rules. In order to extract diverse rules and control the number of rule sets, rules are checked and extracted according to the distance between rules in the same type of rule set and rules in different types of rule sets.

Likewise, Sun et al. [13] proposed a WSN-NSA intrusion detection model based on the improved V-detector algorithm for wireless sensor networks (WSN). The V-detector algorithm is modified by modifying detector generation rules and optimizing detectors, and principal component analysis is used to reduce detection features. Similarly, Tajbakhsh et al. [14] proposed an intrusion detection model based on fuzzy association rules, which uses fuzzy association rules to construct classifiers, and uses some matching metrics to evaluate the compatibility of any new samples with different rule sets.

Singh et al. [15] proposed an advanced hybrid intrusion detection system (AHIDS) that automatically detects wireless sensor network attacks. Moreover, authors in [16] proposed a method of using the synthetic minority oversampling technique (SMOTE) to balance the dataset and then uses the random forest algorithm to train the classifier for intrusion detection. The simulations are conducted on a benchmark intrusion dataset, and the accuracy of the random forest algorithm has reached 92.39%, which is higher than other comparison algorithms.

The rest of this paper is organized as follows. Section 2 illustrates reviews on related works with some background. In Section 3, our proposed Multi-Layer detection model is demonstrated. Then, Section 4 presents the implementation and the experimental results obtained from our proposed model. the results' analysis and discussions were clarified in Section 5. Finally, conclusions and future work are presented in Section 6.

II. RELATED WORKS AND BACKGROUND

WSN faces threats and security issues during the transmission process of data packets between its elements. This is mainly due to the vulnerable nature of WSNs, as these types of network has a considerable number of sensor nodes which are prone to being attacked and receive severe kinds of threats. From the previous studies, we found that such issues have been tackled by abnormal detection methods [17], [18], [19] and misuse detection methods [20], [21]. Authors in [22] proposed an anomaly detection framework in heterogeneous WSNs using real-data. They combined two different approaches: the first approach is the short-term approach, which locally analyzed the data that sense the individual nodes; the second approach is the long-term approach that compares data coming from several heterogeneous sensors over the network. The proposed framework demonstrated a combination of short-long term approaches which can reduce the drawbacks of using each of them separately and gives better performance.

According to [1], the authors presented an intrusion detection method for wireless networks based on improved Conventional Neural Network (ICNN) by first pre-processing the network traffic data, and then used the ICNN to model that data. Their results give an improved accuracy and a higher true positive rate of intrusion detection; it also gives a lower false positive rates compared with the other models. In the work presented by [23], an approach for jamming detection in WSN is proposed based on cooperation with the feedback received from the other connected neighbor's nodes. The model used two techniques, a connected mechanism and an extended mechanism. the results display that this model is more effective when applied on a hierarchical protocol like the Multi-Parent hierarchical.

Another intrusion detection model based on deep learning was proposed by [2]. They built a Deep Belief Network (DBN) combined with multi-restricted Boltzmann machine (RBM), in addition to using the support vector machine (SVM) in training the model. Their experimental results showed that the proposed detection model improved the detection accuracy. An intelligent WSN intrusion detection approach was introduced by [24], which shows that it could decrease the attacks efficiently. They proposed an Artificial Neural Network classifier with Multilayer Perceptron (ANN-MLP) by using holdout and 10-Fold cross-validation methods. In addition to building their own dataset that specialized for the WSN attacks. Their results concluded that with one hidden layer they got the most high accuracy values; however, their approach was mainly based on one detection layer that applies a very computationally expensive learning method.

III. THE PROPOSED MULTI-LAYER DETECTION MODEL

In this paper, we propose a framework for intrusion detection in WSN, that is shield with a defence in depth strategy; leading to an increased security of the working system as a whole. Fig. 1 shows an overview of the system, where the two protection layers represented as the Edge-based Method, and the Cloud-based Method; both layers deploy a machine learning algorithms to facilitate the process of identifying non-previously seen network attacks. This is an extension work of our recent research paper [25]. The following subsections described the deployed methods in details:

A. First Detection Layer: Naive bayes-based Method

In order to avoid complexity and overwhelming the first detection layer, we chose to implement a binary classifier where the traffic is classified to either, normal or malicious traffic only [26], [27]. We have used Naive bayes algorithm as a base of the classifier, due to its simplicity and computational efficiency, that makes it a promising choice for real-time decision making of the inspected packets.

Naive Bayes classifier is based on the well-known Bayesian theorem; and it is particularly suited to high-dimensional datasets [28]. Despite its relative simplicity, in many complex real-world conditions this classifier works very well and it might outperform more sophisticated classification methods. Naive Bayes model allows each attribute to contribute equally and independently to the final decision, in which it results in being more computationally efficient compared to other classifiers.

B. Second Detection Layer: Random Forest-based Method

As discussed in the previous subsection, the first layer will classify the monitored traffic into either: normal or malicious traffic, with no further details in terms of the attack type; this is mainly due to the fact that on that layer we are mainly seeking for simplicity and time efficiency of the decision making process. However, as the second detection layer is located on the cloud and mainly handle the suspicious traffic, there will be less complications in terms of the provided resources, meaning that more complex algorithms and more thorough analysis could be carried out. Therefore, the Random Forest (RF) with multi-class classifier has been used to confirm the traffic with the malicious intent; the classifier has been used also to identify the type of the launched attack, thus, providing guidelines for choosing the appropriate defence mechanism.

Random Forest classifier composed of a set of Decision trees, where every tree provides an insight about each sample's class. At the end of the classification, the class with the most votes is selected as the likely class. The aggregation approach follows in this classifier is based on Breiman 's concept of bagging with randomly selected features on each generated bag, thus creating a set of variation decision trees [29]. Decision trees, which are constructed during the classification task on Random forest classifier, are supervised learning algorithms that are used to address both classification and regression tasks. They originates rules from training several samples represented by a set of attributes; where they derives specific rules that can be easily interpreted as they are visualized as a tree-like graph.

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Python 3.7 has been used to implement the proposed framework, in addition to using the latest version of Scikit-learn, which is an open source machine learning library [24]. For the testing purposes, we have used WSN-DS, which is a dataset generated mainly for intrusion detection systems in wireless sensor networks.

A number of metrics have been utilized to assist and evaluate the performance of the implemented system, those metrics could be described briefly as follows:

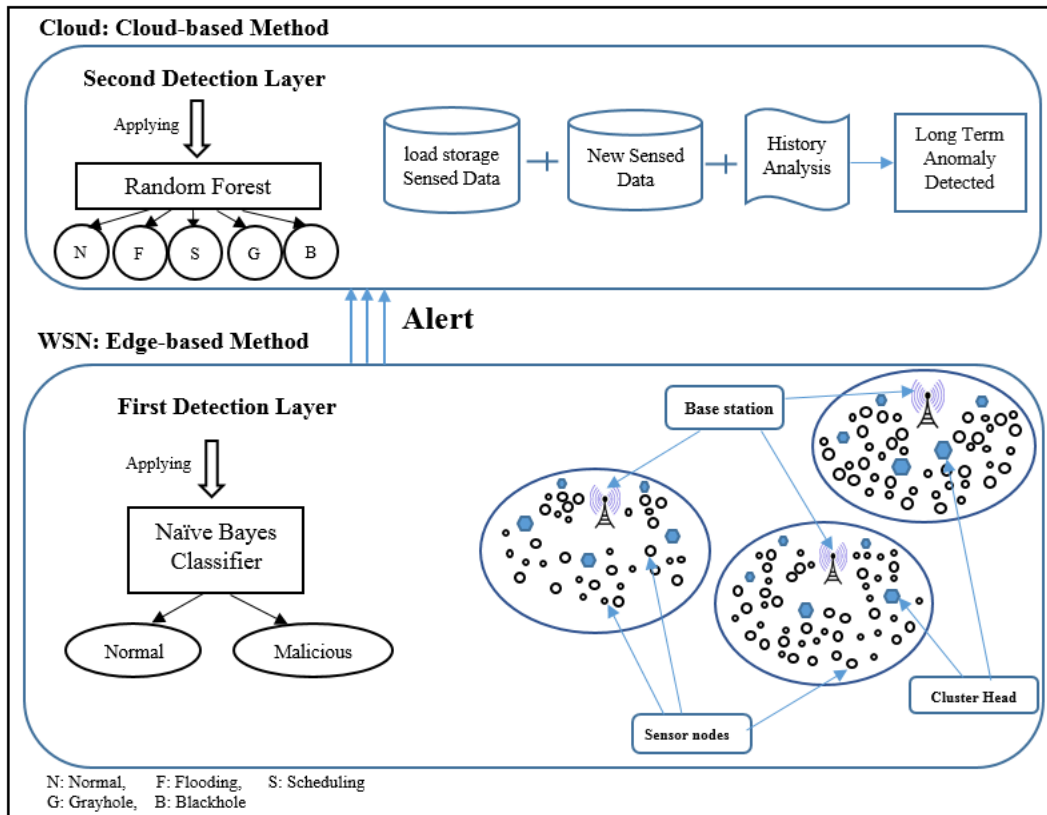


Fig. 1. The Proposed Multi-Layer Detection Model, with Two Protection Layers.

- True positive (TP): the number of network connections correctly identified as attacks.
- True negative (TN): the number of network connections correctly identified as normal connections.
- False positive (FP): the number of network connections incorrectly identified as attacks.
- False negative (FN): the number of network connections incorrectly identified as normal connections.

Those terms have been used to derive different evaluation metrics, i.e. the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR); in addition, they have been used also to calculate the *Precision* (P), as follows:

$$TPR = TP / (TP + FN) \quad (1)$$

$$TNR = TN / (TN + FP) \quad (2)$$

$$FPR = FP / (FP + TN) \quad (3)$$

$$FNR = FN / (FN + TP) \quad (4)$$

$$Precision = TP / (TP + FP) \quad (5)$$

To establish the feasibility of the proposed approach, and to determine its accuracy we have used a dataset generated mainly for evaluating Intrusion Detection Systems in Wireless Sensor Networks (referred to as the WSN-DS) [24]. The dataset consists of a number of 19 features monitored during normal

and abnormal scenarios, where in the latter various number and types of Denial of Service (DOS) attacks were simulated (i.e. Blackhole, Grayhole, Flooding, and Scheduling attacks (TDMA)). Table I gives an overall view of the WSN-DS dataset features including their description.

V. RESULTS ANALYSIS AND DISCUSSIONS

A. First-Layer Results and Discussions

As the main purpose of the first layer is identifying the abnormal traffic with the least resources possible, we used Mutual information (MI) algorithm to quantify the importance of each feature (as seen in Fig. 2), therefore, selecting the most relevant ones; MI is widely known as a good indicator to determine the relevance between variables, and it is usually used in the area of AI as a feature selection algorithm [30], [31]. Fig. 2 emphasises the computed MI score for each feature, where the higher the score, the more important the feature.

Based on some preliminary tests, we have found that choosing the best three features, as ranked by MI, will give the highest classification performance. Fig. 3 (a & b) shows the classification accuracy when including the best three features, and all of the 19 features provided by WSN-DS, respectively. Thus, the first three features, ADV_S, Is_CH, and Join_S, have been used as an input to the Naive Bayes classifier in order to filter the malicious traffic and pass it to the second protection layer for further examination. It can be seen from Fig. 3 (a) that a 99% detection accuracy of the abnormal activities has

TABLE I. THE WSN-DS DATASET FEATURES DESCRIPTION

No.	Feature	Description
1	id	unique ID to distinguish the sensor node
2	Time	current node simulation time
3	Is_CH	distinguish whether the node is Cluster Head
4	who CH	ID of the CH in the current round
5	Dist_To_CH	distance between the node and its CH
6	ADV_S	number of advertised CH sent messages
7	ADV_R	number of advertised CH received messages
8	JOIN_S	number of joined request CH sent
9	JOIN_R	number of joined request CH received
10	SCH_S	number of scheduled CH sent messages
11	SCH_R	number of scheduled CH received messages
12	Rank	order of this node in the schedule
13	DATA_S	number of data packets sent to CH
14	DATA_R	number of data packets received from CH
15	Data_Sent_To_BS	number of data packets sent to BaseStation
16	dist_CH_To_BS	distance between CH and BS
17	send_code	the cluster sending code
18	Consumed Energy	the amount of energy consumed in the round
19	Attack type	the type of the attack

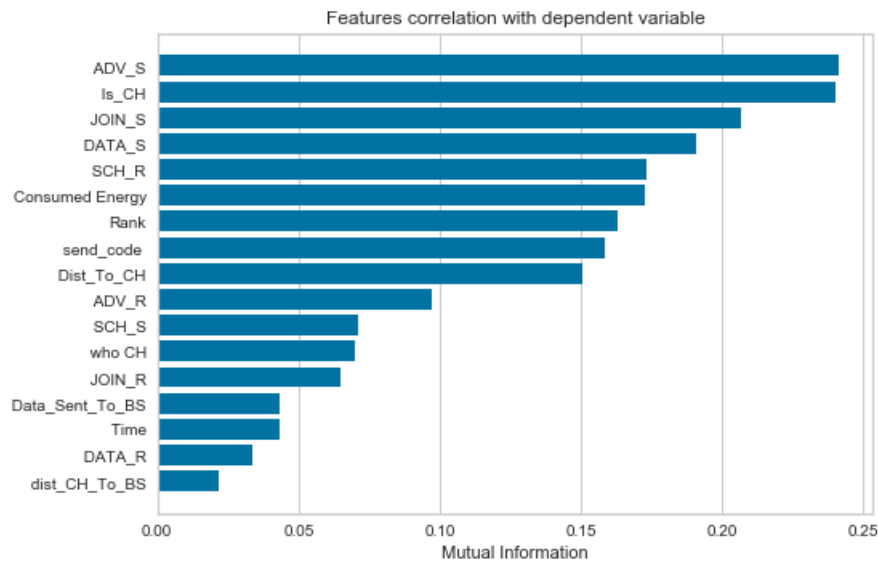


Fig. 2. MI Score for Each Monitored Feature.

be achieved with the use of 3 features only, while maintaining a low usage of computational resources; the Area Under the Curve (AUC), which is a commonly used stat to show the overall performance of a classification method, is also shown on Fig. 4.

B. Second-Layer Results and Discussions

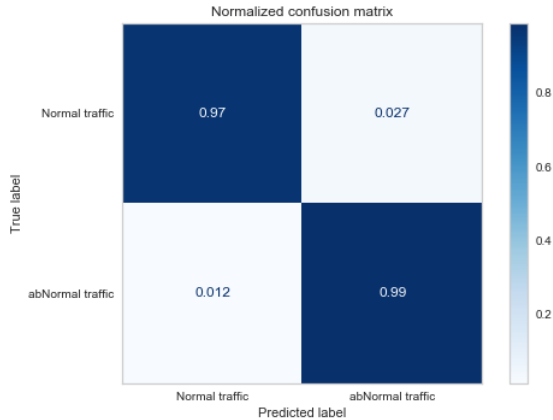
On the second detection layer, more examination of the malicious traffic will be carried out; thus, a multi-class classification using RF classifier is performed to identify the specific type of the attack, thereby choosing the appropriate defence mechanism. Classification results obtained by RF classifier is shown on Fig. 5; it could be seen that a relatively high performance was achieved as illustrated in Table II. Therefore, such a high detection performance allows more concrete countermeasures to be adopted automatically by the system.

Generally, the aim of an IDS is to obtain a high precision [32], as this measure shows how many cases, predicted as an intrusive, are actually correct. Based on that, when we compare the performance obtained with the RF classifier in this paper with a previous work that used the same dataset, e.g. [24], it could be clearly seen that a higher precision has been achieved, where the precision of the attacks detection were 73%, 90%, 99.5%, 91.1%, and 99% in Blackhole, Flooding, Scheduling, and Grayhole attacks, in addition to the normal case (without attacks), respectively. A comparison of the performance metrics between the previous work done by [24] and our proposed model is illustrated in Fig. 6 & 7, which show an improvement in the performance values of TPR, TNR, FPR, FNR and Precision, compared to the previous work.

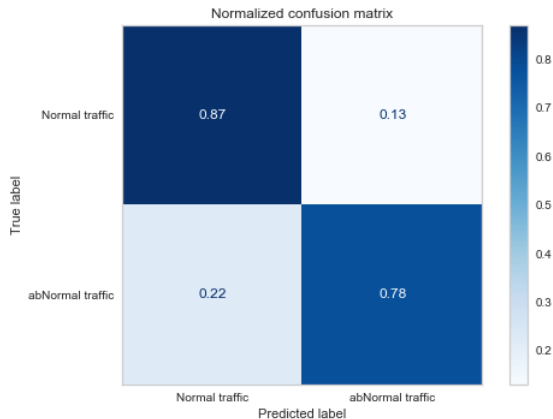
However, Fig. 6 shows one case where our proposed work has achieved a slightly lower value; this is the case of the TNR

TABLE II. RF PERFORMANCE OF 10-FOLD CROSS-VALIDATION COMPARED WITH THE PREVIOUS WORK

Attack Type	The previous work results					The proposed results					P % Change
	TPR	FPR	FNR	TNR	P	TPR	FPR	FNR	TNR	P	
Normal	0.998	0.018	0.002	0.982	0.998	0.998	0.023	0.002	0.977	1.0	+0.2%
Flooding	0.994	0.001	0.006	0.999	0.904	0.991	0.001	0.009	0.999	0.904	0
Scheduling	0.922	0	0.078	1.0	0.995	0.927	0.0	0.073	1.0	0.995	0
Grayhole	0.756	0.003	0.244	0.997	0.911	0.955	0.001	0.045	0.999	0.970	+6.5%
Blackhole	0.928	0.009	0.072	0.991	0.730	0.991	0.001	0.009	0.999	0.999	+37%



(a) With including only the 3 best features selected based on MI



(b) With including all the features

Fig. 3. Classifying Network Attacks using Naive bayes Classifier.

of the Normal packets. Consequently, the FPR derived from the Normal packets becomes higher. In such a case, this means that more packets will be inspected further, and flagged as malicious, although they do not carry any harmful intentions. This case could be costly (in terms of the time spent during the investigation); however, it would not be as expensive as if a malicious packet has been missed to be identified, and instead recognised as a Normal one.

Moreover, our work provides other advantages inherited by the use of RF classifier (rather than artificial neural network on [24]), such as the fact that it is considered less computationally expensive compared with ANN classifier. The usage of RF classifier also increases the performance of the security of the

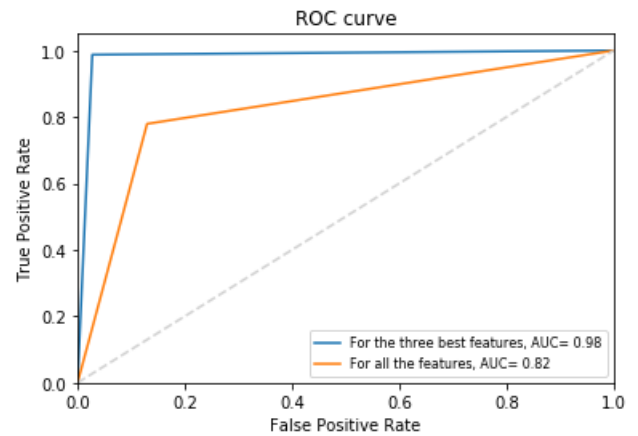


Fig. 4. ROC Curve Showed Comparison between the Classification Results for All the Features and the Three best Only.

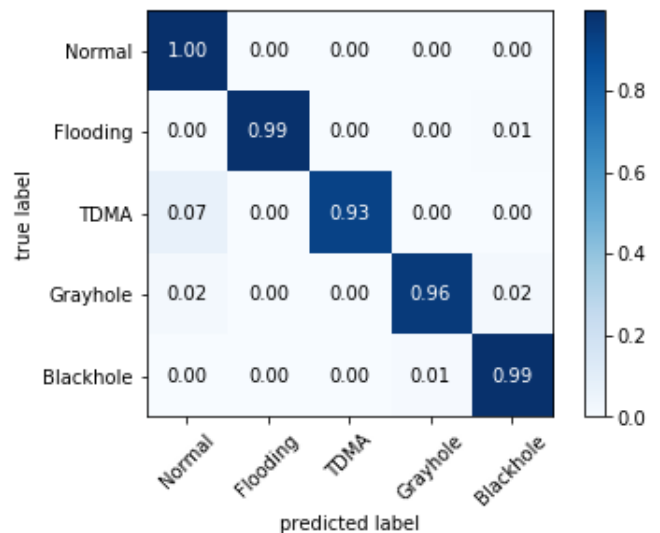


Fig. 5. Classifying Network Attacks using Random Forest Classifier (Performance Rounded to Two Decimal Points).

system as a whole in that it provides the interpretability and transparency of the results, as shown in Fig. 8 where the result of a tree generated by RF classifier could be easily interpreted; the resulting rules could also be investigated further using tools such as [33]. Such properties are very important in the analysis of the attacks, optimisation and handling of the system errors [34]. Most importantly, the proposed work employs a layered defence mechanism that enhances the security by providing

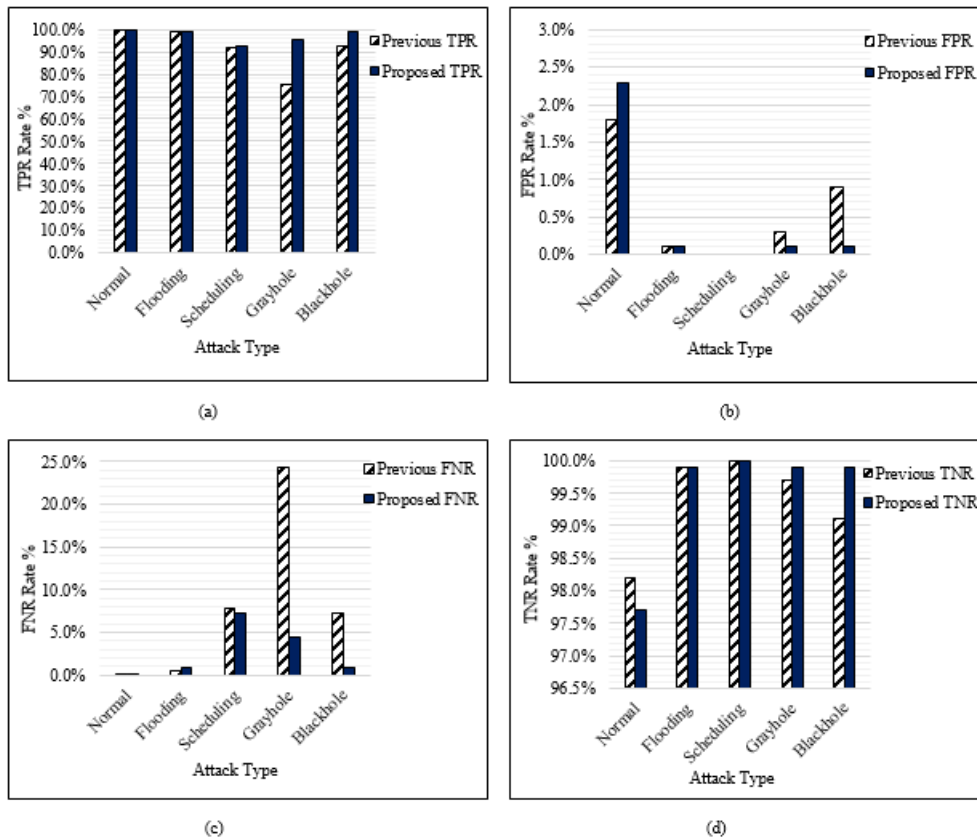


Fig. 6. Performance Metrics Results for the Previous the Proposed Multi-layer Model; (a) TPR, (b) FPR, (c) FNR, and (d) TNR.

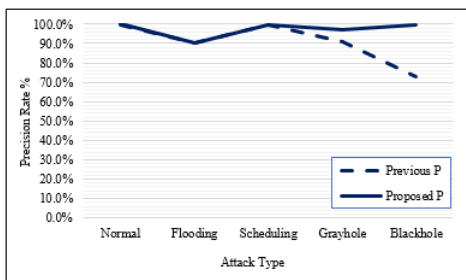


Fig. 7. Precision Improvement for the Proposed Multi-layer Model.

an extra protection layer to defend the whole system in cases where the first layer has been bypassed or fail as a result to the ever-changing attack techniques, and the present increasing threat landscape.

VI. CONCLUSIONS

Intrusion detection in wireless sensor networks is a very challenging task. The majority of the current WSN intrusion detection models were using machine learning methods, but they apply only one method for the whole network. In this paper, we propose a multi-layer framework for intrusion detection system in WSN, leading to increase the network security. Our proposed model consists of two consequent protection layers; the first layer is located on the edge of the network where the sensors are located. It used the Naive bayes classifier

where the traffic is classified into normal or malicious traffic which achieving simplicity and time efficiency of the decision-making process. While the second layer is located on the cloud, and mainly handle the suspicious traffic by using a multi-class Random Forest classifier.

The implementation results demonstrate that our proposed multi-layer protection model improved the values of TPR, TNR, FPR, and FNR in addition to achieving a high Precision rate with values 100%, 90.4%, 99.5%, 97%, 99.9% for the Normal, Flooding, Scheduling, Grayhole, and Blackhole attacks, respectively. While the previous work has the values 99.8%, 90.4%, 99.5%, 91.1%, 73% for the Normal, Flooding, Scheduling, Grayhole, and Blackhole attacks, respectively. Nevertheless, the results in Fig. 6 show only one case where our proposed work has achieved a slightly lower value; this is the case of the TNR of the Normal packets. Consequently, the FPR derived from the Normal packets becomes higher. In such an instance, this means that more packets will be inspected further, and flagged as malicious, although they do not carry any harmful intentions. This case could be costly (in terms of the investigation time); however, it would not be as expensive as if a malicious packet has been missed to be identified, and instead recognised as a Normal one.

As future work, we plan to improve the performance of our multi-layer detection model in WSN by using one of the deep learning techniques in the second layer, where the higher number of attacks types appear.

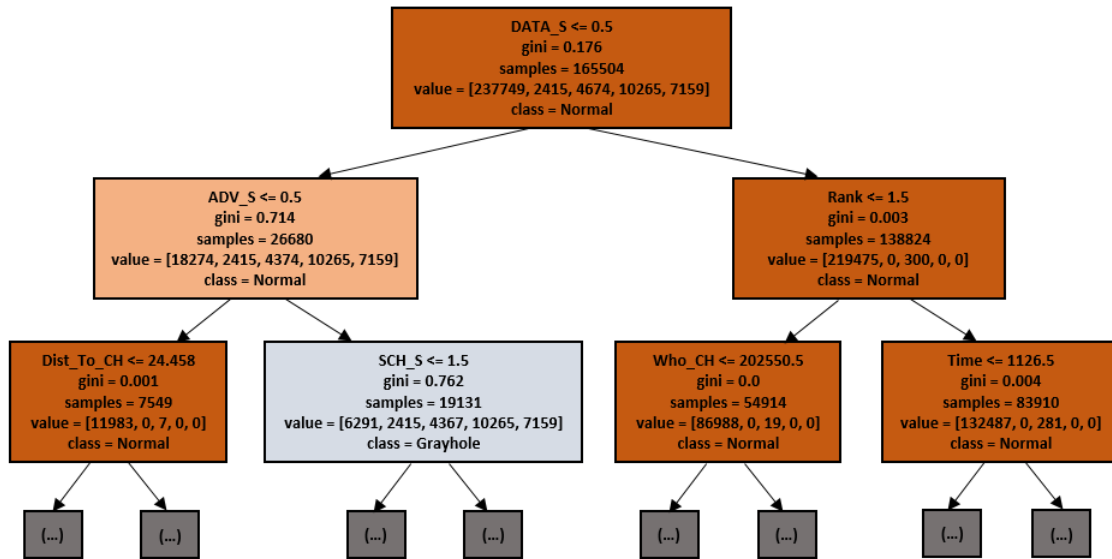


Fig. 8. A sample of the Tress Generated by the Random Forest Classifier, where Interpretable and Transparent Rules are Obtained During the Classification Task.

REFERENCES

- [1] H. Yang and F. Wang, "Wireless network intrusion detection based on improved convolutional neural network," *IEEE Access*, vol. 7, pp. 64366–64374, 2019.
- [2] H. Yang, G. Qin, and L. Ye, "Combined Wireless Network Intrusion Detection Model Based on Deep Learning," *IEEE Access*, vol. 7, pp. 82624–82632, 2019.
- [3] S.S. Roy, A. Mallik, R. Gulati, M.S. Obaidat, P.V. Krishna, "A deep learning based artificial neural network approach for intrusion detection," In *International Conference on Mathematics and Computing*, Springer, Singapore, pp. 44–53, 2017.
- [4] Y. Liu, S. Liu, and X. Zhao, "Intrusion detection algorithm based on convolutional neural network. *Transactions on Engineering and Technology Research*, vol. 37, pp. 1271–1275, 2019.
- [5] M. Wang, and J. Liu, "Network intrusion detection based on convolutional neural network," *Net information Security*, vol. 3, pp. 990–994, 2019.
- [6] M.A. Yong, "A network intrusion detection schemer based on fuzzy inference and Michigan genetic algorithm," *Electron. Des. Eng.*, vol. 24, pp. 107–110, 2016.
- [7] Q. Yaun, and L.T. Lv, "Network intrusion detection method based on combination of improved ant colony optimization and genetic algorithm," *J. Chongqing Univ. Posts Telecommun*, vol. 29, pp. 85–89, 2019.
- [8] H. Chen, G.X. Wan, Z.J. Xiao, "Intrusion detection method of deep belief network model based on optimization of data processing," *Journal of Computer Applications*, vol. 37, pp. 1636–1643, 2017.
- [9] C. Yin, Y. Zhu, J. Fei, X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21594–21961, 2017.
- [10] N. Shone, T.N. Ngoc, V.D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, pp. 41–50, 2018.
- [11] C. Giuseppe, C. Calafiore, and F. Giulia, "Sparse ℓ_1 and ℓ_2 Center Classifiers," *arXiv preprint arXiv:1911.07320*, 2019.
- [12] N. Lu, Y. Sun, H. Liu, and S. Li, "Intrusion Detection System Based on Evolving Rules for Wireless Sensor Networks," *J. Sens.*, vol. 2018, pp. 1–8, 2018.
- [13] Z. Sun, Y. Xu, G. Liang, and Z. Zhou, "An Intrusion Detection Model for Wireless Sensor Networks with an Improved V-Detector Algorithm," *IEEE Sens. J.*, vol. 18, pp. 1971–1984, 2018.
- [14] A. Tajbakhsh, M. Rahmati, and A. Mirzaei, "Intrusion detection using fuzzy association rules," *Appl. Soft. Comput.*, vol. 9, pp. 462–469, 2009.
- [15] R. Singh, J. Singh, and R. Singh, "Fuzzy Based Advanced Hybrid Intrusion Detection System to Detect Malicious Nodes in Wireless Sensor Networks," *Wirel. Commun. Mob. Comput.*, vol. 2017, pp. 1–14, 2017.
- [16] X. Tan, S. Su, Z. Huang, X. Guo, Z. Zuo, X. Sun, and L. Li, "Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm," *Sensors*, vol. 19, no. 1, pp.203–218, 2019.
- [17] P. Li, W.H. Zhou, "Hybrid intrusion detection algorithm based on k-means and decision tree," *Computer Modernization*, vol. 37, pp. 12–16, 2019.
- [18] Y.M. Qi, L. Ming, F. Yanming, "Research on SVM network intrusion detection based on PCA," *Information Network Security*, vol. 2, pp. 15–18, 2015.
- [19] X. Wang, "Design of temporal sequence association rule based intrusion detection behavior detection system for distributed network," *Modern Electron. Techn.*, vol. 41, pp. 108–114, 2018.
- [20] K. Zheng, Z. Cai, X. Zhang, Z. Wang, and B. Yang, "Algorithms to speedup pattern matching for network intrusion detection systems," *Computer communications*, vol. 62, pp. 47–58, 2015.
- [21] S. Kim, "Pattern matching acceleration for network intrusion detection systems. In *International Workshop on Embedded Computer Systems*. Springer, Berlin, Heidelberg, 2005; pp. 289–298, 2005.
- [22] F. Cauteruccio, G. Fortino, A. Guerrieri, A. Liotta, D.C. Mocanu, C. Perra, G. Terracina, and M.T. Vega, "Short-long term anomaly detection in wireless sensor networks based on machine learning and multi-parameterized edit distance," *Information Fusion*, vol. 52, pp. 13–30, 2019.
- [23] C.; Del-Valle-Soto, L.J.; Valdivia, and J.C. Rosas-Caro, "Novel detection methods for securing wireless sensor network performance under intrusion jamming," In the *International Conference on Electronics, Communications and Computers (CONIELECOMP)*, IEEE, pp. 1–8, 2019.
- [24] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, "WSN-DS: A dataset for intrusion detection systems in wireless sensor networks," *Journal of Sensors; Hindawi*, vol. 2016, pp. 1–16, 2016.
- [25] D.M.; Ibrahim, and N.M. Alruhaily, "Anomaly detection in Wireless Sensor Networks: A Proposed Framework," *International Journal of Interactive Mobile Technologies (IJIM)*, vol. 14, pp. 150–158, 2020.
- [26] S.L. Ting, W.H. Ip, Tsang, and H.C. Albert, "Is Naive Bayes a good classifier for document classification," *International Journal of Software Engineering and Its Applications*, vol. 5, pp. 37–46, 2011.

- [27] E. Frank, Bouckaert, and R. Remco, "Naive bayes for text classification with unbalanced classes," In European Conference on Principles of Data Mining and Knowledge Discovery, Springer, pp. 503–510, 2006.
- [28] A. El Abdouli, L. Hassouni, and H. Anoun, "Sentiment Analysis of Moroccan Tweets using Naive Bayes Algorithm," International Journal of Computer Science and Information Security (IJCSIS), vol. 15, 2007.
- [29] L. Breiman, "Bagging predictors Machine learning," Springer, vol. 24, pp. 123–140, 1996.
- [30] N. Kwak, and C.H. Choi, "Feature selection by mutual information based on Parzen window," IEEE transactions on pattern analysis and machine intelligence, vol. 24, pp. 1667–1671, 2002.
- [31] N. Barraza, S. Moro, M. Ferreyra, and A. de la Peña, "Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study," Journal of Information Science, vol. 45, pp. 53–67, 2019.
- [32] Ghorbani, A.A.; Lu, W. Tavallae, M. Network intrusion detection and prevention: concepts and techniques. In Springer Science & Business Media, 2009.
- [33] Ribeiro, M.T., Singh, S. and Guestrin, C. Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, 2016.
- [34] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectively: An evaluation of neural networks, random forest, regression trees and support vector machines," Ore Geology Reviews, Elsevier, vol. 71, pp. 804–818, 2015.

ParaDist-HMM: A Parallel Distributed Implementation of Hidden Markov Model for Big Data Analytics using Spark

Imad Sassi^{*1}, Samir Anter², Abdelkrim Bekkhoucha³
Computer Science Laboratory (LIM), FSTM, Hassan II University,
Casablanca, Morocco

Abstract—Big Data is an extremely massive amount of heterogeneous and multisource data which often requires fast processing and real time analysis. Solving big data analytics problems needs powerful platforms to handle this enormous mass of data and efficient machine learning algorithms to allow the use of big data full potential. Hidden Markov models are statistical models, rich and widely used in various fields especially for time varying data sequences modeling and analysis. They owe their success to the existence of many efficient and reliable algorithms. In this paper, we present ParaDist-HMM, a parallel distributed implementation of hidden Markov model for modeling and solving big data analytics problems. We describe the development and the implementation of the improved algorithms and we propose a Spark-based approach consisting in a parallel distributed big data architecture in cloud computing environment, to put the proposed algorithms into practice. We evaluated the model on synthetic and real financial data in terms of running time, speedup and prediction quality which is measured by using the accuracy and the root mean square error. Experimental results demonstrate that ParaDist-HMM algorithms outperforms other implementations of hidden Markov models in terms of processing speed, accuracy and therefore in efficiency and effectiveness.

Keywords—Big data; machine learning; Hidden Markov model; forward; backward; baum-welch; parallel distributed computing; spark; cloud computing; ParaDist-HMM

I. INTRODUCTION

Big data is an extremely large, typically heterogeneous, structured and unstructured data, gathered from a wide range of sources (logs files, Internet of Things [1], web, transactions, social media insights, sensors, mobile devices, third party data, etc.), with a very high speed of generation and diffusion which often requires fast processing and real time analysis [2].

Everyday, huge volume of data is produced in different fields, such as commerce, medicine, social media, or Internet of Things which is compiling data in an accelerated way. So, how can we succeed to draw valuable insights from these data?

The characteristics of big data (volume, velocity and variety) have given rise to numerous challenges in the domain of big data analytics, for instance, scalability of models, efficiency of algorithms and robustness of hardware configurations [4].

Regarding the volume of data, classical solutions, which use traditional data warehouses, are limited because their latency is too long and the data must first be stored in single place, which is not recommended for the security of critical data for example [8].

The velocity is also a key factor for data analysis efficiency. Usually, the data has to be processed in a very short time, even in real time, so that we get the good information in good time. Thus, big data analysis requires powerful algorithms in order to make all of this data very quickly understandable and to use it effectively in decision making in a constantly evolving environment. Computing power and speed of analysis are therefore essential [9].

The diversity and complexity of data formats are also causing real problems since data is collected from various sources. Faced with this challenge, classical algorithms have to be ameliorated in order to manage the variety of data [10].

In addition, the big data universe is undergoing great technological evolution. Spark [11], Hadoop [12], graph analytic [13] and GPU distributed computing are now ubiquitous solutions in many sectors.

Given the above, the use of the full potential of big data will be achieved by efficient processing that requires new techniques and algorithms referred to big data analytics or data science. Among these techniques, machine learning whose objective is to create systems that can learn from the data they receive. This principle of machine learning explains its renewed interest with the appearance of big data since this enormous amount of knowledge-bearing data and this computation power makes it possible to manage more and more data and thus, to refine the relevance of predictions of learning systems [3].

Numerous studies have shown that many factors can affect the implementation efficiency of algorithms for big data analytics. Among these factors the computation time, the memory cost, the hardware architecture, the scalability and centralization, the non-dynamic of most traditional data analysis methods, the analyze of social network data, the security and privacy issues. Thus, several problems arise when handling and analyzing big data [5–7].

Solving these problems will contribute to facilitating knowledge discovery and decision making and it will undoubtedly open new perspectives for researchers in the field of big data analytics, and this will influence positively the global growth and will contribute to the development of business strategies and models in several sectors.

To achieve this goal, new flexible big data analytics solutions are needed. In this context, the parallel distributed

computing approach, which has brilliantly succeeded in the past decade, is one of the most promising solutions [14].

It is one of the efficient analysis methods that have shown their excellent performance in this type of application. Given the importance of emerging big data technologies it has now become a requirement to use them for implementing parallel distributed computing. However, there are great challenges regarding the design of parallel distributed implementations, related to algorithms and frameworks, mainly, the communication errors, the storage and the query burden and the integration of massive heterogeneous big data into a single unified view, the matrix multiplications and the optimization techniques [15–17].

The combination of classical algorithms and big data technologies enables a high level of flexibility, allows the simultaneous execution of several complex analyzes, and facilitates the integration of new analysis tools.

One of the most powerful machine learning algorithms are hidden Markov models (HMMs) [18]. HMMs are widely used for sequential data modeling and time series analysis. They owe their success to the existence of many efficient and reliable algorithms. Given the great potential demonstrated by the paradigm of HMMs in various applications, it seems quite natural to extend them for big data. Although there are many parallel implementations for HMMs, there is no clear compromise for each application scenario, especially for real-time processing of large data of different structures.

To address some of the aforementioned issues, this paper presents a new Spark-based parallel distributed implementation of HMMs to make their use for modeling and analysis applicable for big data without decreasing in accuracy and computational efficiency. Our aim is to provide a solution for big data analytics that meets two fundamental criteria for designing big data solutions: an architectural criterion (an architecture that supports parallel computations and distributed storage) and an algorithmic criterion (algorithms capable of efficiently processing and analyzing big data).

In summary, the main contributions of this work are:

- We introduce the phenomenon of big data and we explain the need for new machine learning algorithms to draw value from this huge amount of data.
- We present a detailed study of hidden Markov models and we describe its three fundamental problems (evaluation, decoding and training).
- We review the existing solutions with a description and analysis of the main parallel implementations of hidden Markov model algorithms.
- We propose new parallel distributed versions of the Forward, Backward and Baum-Welch algorithms, then we describe a proposed Spark-based big data architecture to use the new algorithms.
- We experimentally evaluate the proposed algorithms in a cloud computing environment using a set of synthetic and real-world data, and we compare the performances of these algorithms with classical ones, but also with the main solutions proposed in the benchmark.

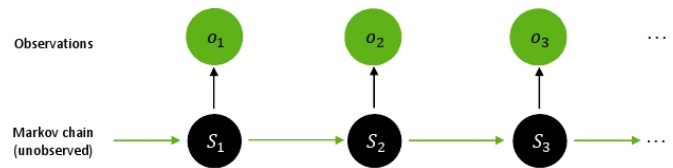


Fig. 1. Basic Structure of a Hidden Markov Model.

The rest of the paper is organized as follows. Section 2 gives a formal study of hidden Markov models, discusses main parallel distributed implementations challenges and reviews some proposed solutions for parallelism of HMMs algorithms. Section 3 describes the studied problem and shows the novelty of this research. In Section 4, we present main concepts of the proposed approach, then we describe the new parallel distributed HMM algorithms (ParaDist-HMM) and the proposed big data architecture to put them into practice. Section 5 presents the experiments settings and methods used for the evaluation of the algorithms. The results of the experimental study are presented and discussed in Section 6. Finally, Section 7 draws the conclusions of the paper and gives some prospective points for the future work of this research.

II. BACKGROUND AND RELATED WORK

In this section, first, we provide an overview of the theoretical and technical background required for this study. Next, we discuss the fundamental challenges of parallel distributed implementations of machine learning algorithms in the era of big data. Then, we present main related works with a study of main advantages and limitations of these works.

A. Hidden Markov Models

In the literature, there is a large amount of studies of HMMs [18–20]. Based on these interesting studies, in this section, we will present the theoretical foundations of the HMMs, in particular, the algorithms studied in this article.

There are different definitions for HMMs. One of the most well-known definitions in the literature is provided by Rabiner and Juang [21] who define a HMM as a “doubly stochastic process with an unobservable underlying stochastic process (hidden), but can only be observed by another set of stochastic processes that produce the sequence of observed symbols”. It consists of two stochastic processes. The first is a Markov chain characterized by states and transition probabilities where the states of the chain are not visible, so “hidden”. The second produces emissions observable at each instant based on a state-dependent probability distribution. Thus, we can simply analyze what we observe without seeing at which states it occurred. The observations can be discrete or continuous. It is important to note that the “hidden” denomination of a HMM refers to the states of the Markov chain and not to the model parameters (see Fig. 1). In the rest of this section, we will present the essential notation and key concepts about HMMs which will be helpful in the rest of this work.

In order to fully define a HMM, the following elements must be defined:

1. The N states of the model, defined by

$$S = \{S_1, \dots, S_N\}$$

2. The M observation symbols per state $V = \{v_1, \dots, v_M\}$ corresponding to the output of the system being modeled. If they are continuous then M is infinite.

3. The state transition probability distribution $A = \{a_{ij}\}$, where a_{ij} is the probability that the state at time $t + 1$ is S_j given that the state at time t is S_i .

$$a_{ij} = Pr\{q_{t+1} = S_j \mid q_t = S_i\}, 1 \leq i, j \leq N \quad (1)$$

The transition probabilities must satisfy the normal stochastic constraints:

$$a_{ij} \geq 0, 1 \leq i, j \leq N \text{ and } \sum_{j=1}^N a_{ij} = 1, 1 \leq i \leq N \quad (2)$$

4. The observation symbol probability distribution in each state, $B = \{b_j(v_k)\}$ where $b_j(v_k)$ is the probability that symbol v_k is emitted in state S_j .

$$b_j(v_k) = Pr\{o_t = v_k \mid q_t = S_j\}, 1 \leq j \leq N, 1 \leq k \leq M \quad (3)$$

where v_k denotes the k^{th} observation symbol in the alphabet and o_t the current parameter vector. The observation may be discrete or continuous.

The following stochastic constraints must be satisfied:

$$b_j(v_k) \geq 0 \text{ and } \sum_{k=1}^M b_j(v_k) = 1, 1 \leq j \leq N, 1 \leq k \leq M \quad (4)$$

5. The HMM is the initial state probability distribution $\Pi = \{\pi_i\}$, where π_i is the probability that the model is in state S_i at the time $t = 0$ with

$$\pi_i = Pr\{q_1 = S_i\}, 1 \leq i \leq N, \sum_{i=1}^N \pi_i = 1 \quad (5)$$

The following notation $\lambda = (A, B, \Pi)$ is often used in the literature to denote a discrete HMM.

We will also use the notations $Pr\{O|\lambda\}$: the probability that the given observations $O = o_1, o_2, \dots, o_T$ are generated by a model λ with a given HMM. $\alpha_t(i)$: the forward variable is the probability of the partial observation sequence o_1, o_2, \dots, o_t to be produced by all possible state sequences that end at i^{th} state and that we are in state S_i at time t . $\beta_t(i)$: the backward variable is the probability of the partial observation sequence $o_{t+1}, o_{t+2}, \dots, o_T$ given that the current state is S_i . $\gamma_t(i)$: the probability of being at state S_i at time t , given the model λ and the observation O and $\xi_t(i, j)$: the probability of being at state S_i at time t and at state S_j at time $t + 1$, given the model λ and the observation O .

There are three fundamental problems studied around HMMs. First, the evaluation problem in which we try to calculate the probability $Pr\{O|\lambda\}$ that a given observations O are generated by a model λ with a given HMM. The methods commonly used to solve this problem are the forward or the backward algorithms based on the technique of dynamic programming. Second, the decoding problem in which, we look for the most likely state sequence in a given model λ that produced a given observations O . Viterbi algorithm is the most used to solve this problem [22]. Third, the learning problem in

which we try to adjust the parameters of the model (A, B, Π) to maximize the probability $Pr\{O|\lambda\}$ given a model λ and a sequence of observations O . For this problem, Baum-Welch algorithm (BW), also known as forward-backward algorithm is the most used [19].

In the rest of this article, we focus mainly on the evaluation and the learning problems.

B. Parallel Distributed Implementation Challenges

There is a vast amount of literature concerning challenges to face when designing a parallel distributed implementation. The following table (Table I) presents the most important challenges and criteria, related to the implemented architecture but also to the algorithms in question, to take into account when designing parallel distributed implementations.

C. Related Work

Many practical problems arise during the parallel GPU or CPU implementation of forward, backward, Viterbi or Baum-Welch algorithms for HMMs. This section surveys the solutions proposed in the previous major work on parallel distributed implementation of HMMs. For example, [31], proposes a new distributed multidimensional HMM (DHMM) for multi-object trajectory interaction modeling, the results show superior performance and greater accuracy of the proposed distributed 2D HMM. In [32], the authors present a parallelized HMM to accelerate isolated words speech recognition. Another work of [33] presents a GPU implementation in which they proposed a C and Cuda implementation for the forward, Viterbi and BW algorithms. For a low number of states, the GPU performs far worse than the CPU and for a number of symbols and number of observations, it has had little impact on the difference in speed of execution between the CPU and the GPU. Regarding the execution time, the speed increases can reach 180x for the forward algorithm, 65x for the BW algorithm and 4x for the Viterbi algorithm with 4000 states. In [34] and [35] a proposed C++ library for general HMMs was presented, exploiting modern CPUs with multiple cores and supporting the SSE instruction set to increase performance by distributing the computations for each state among the available processors. The results showed significant accelerations for all conventional HMM algorithms except posterior decoding for a very large number of states. Another parallelization approach has been also proposed for HMMs with small number of states. [36] propose a parallel implementation of the three fundamental algorithms of HMM for GPU computing environment. [37] presented GPU Cuda using Cuda C language and ANSI C language. The result obtained shows an acceleration of the forward-backward implementation faster 4 to 25 times than the classical one. Finally, the work of [38] presented a parallel implementation of a HMM (forward, backward and Viterbi) for the spoken language recognition on the MasPar MP-1. A complexity comparison of the serial and the parallel implementations of the forward and Viterbi algorithms shows that there is a big improvement in execution time.

III. CONTRIBUTION OF THIS WORK

To make big data valuable, we often use machine learning algorithms like HMMs. However, to be efficient in the big

TABLE I. CHALLENGES OF PARALLEL DISTRIBUTED IMPLEMENTATIONS

Authors and Reference	Challenges
Slavakis et al. [23]	communication errors, privacy, incomplete data, storage and query burden, decentralized learning with parallelized multicores, storage in the cloud or using distributed data systems.
Alshamrani et al. [24]	integration of massive heterogeneous big data residing on different sites with different types and formats into a single unified view before starting data mining processes.
Hassan et al. [25]	distributed data mining and multi-agent data extraction since in a distributed environment, traditional techniques require that distributed data be first collected in a data warehouse and pose data confidentiality and sensitivity issues in addition to the costs of storage, communication and computation.
Zhan et al. [16]	matrix multiplication task, the improvement of parallelization of a series of matrix multiplications, parallel programming for shared memory architectures.
Liu et al. [15]	speed up synchronous parallelization, effect of parallelization mechanisms on the overall convergence rate especially when several different techniques are simultaneously used in one machine learning algorithm.
Li et al. [26]	to balance the need of flexibility and generality of machine learning algorithms and the simplicity of systems design.
Zhou et al. [27]	the effect of preprocessing and data probing operations on the efficiency of parallelization, data privacy, inconsistency and skewness issues.
Gunjan et al. [28]	look for new powerful techniques especially divide-and-conquer approaches to decompose problems into several sub problems.
Bhattacharya [29]	rethink optimization techniques used in machine learning algorithms especially with the new requirements of complexity, size and variety of data.
Russell et al. [30]	to think of new advances in logic, in computation, to re-study the theory of probability and to put forward the Neuroscience.

data context, it is necessary to improve the performance of HMMs without losing the quality of the prediction. Through this paper, we aim to provide a parallel distributed implementation of HMMs (i.e., ParaDist-HMM) which ameliorates the performances of previous parallel HMM solutions mainly in terms of execution time, speedup, scalability and accuracy. We also present a big data architecture with horizontal scaling capabilities to manage large volume of both real-time and

batch-based information, based on Spark as a core element which allow to exploit the advantages of its modules for the collection and the storage of heterogeneous data in batch and in real time modes, for data preprocessing (cleaning, extracting, transforming and selecting features) and also for models testing and evaluation. In order to boost processing speeds and to deal with the storage problem, we use cloud platform service which makes available several machines to provide services such as computing and storage.

The parallel distributed computing approach have been chosen for the following reasons:

- On the one hand, to accelerate the performance of classic machine learning algorithms, it is recommended to use a distributed system to speed up analytical tasks. This technique is widely used to manipulate a large amount of data. This is a very efficient technique that ensures data consistency and availability.
- On the other hand, for complex processing, it becomes expensive to maintain analysis requests on a single node due to time latency and hardware requirements. To deal with this problem, the parallelism technique can provide promising solutions. This technique consists in processing data simultaneously, thus making it possible to carry out the greatest number of operations in the shortest possible time.
- Finally, the combination of big data technologies and conventional machine learning algorithms provides a powerful tool to very quickly obtain an overview from huge volumes of unstructured data.

Among the arguments of the proposed approach and the proposed architecture:

- to speed up the learning and prediction process compared to the solutions previously presented and improve the accuracy of the model or at least present performance comparable to previous solutions.
- to offer high scalability of the model.
- it is based, in its implementation, on the distribution of data matrices on several vectors on different nodes unlike the other solutions.
- to handle discrete, continuous and semi-continuous HMMs.
- it can easily be integrated into a big data framework.
- for the computational time consideration, Spark transformation and action reduces the time complexity. The Spark's MLlib library ensures that the quality of the model is not reduced while maintaining much shorter computation times compared to traditional approaches.
- for the calculation time consideration, using a much faster data analysis environment such as Spark reduces the time complexity.
- finally, the power of HMMs offers the possibility of using the model in several application fields.

IV. PROPOSED APPROACH

In this section, firstly, we provide an overview of the Spark's main concepts used to achieve this implementation. Next, we present the proposed approach and we formally define the model and introduce the assumptions and notations. Finally, we provide a description of the big data architecture to put the model into practice for successful big data processing and analysis.

A. Main Spark Concepts used in Parallel Distributed Implementation of HMM

To achieve the implementation of the proposed algorithms we exploited fundamental Spark concepts such as:

1) The use of Resilient Distributed Datasets (RDDs) [11] to split and distribute data into several blocks (See Figure 2a). Since matrices are often quadratically larger than vectors, a reasonable assumption is that vectors fit in memory on a single machine while matrices do not [39, 40]. So, we distribute large matrices over many vectors in several nodes. We used vectors to store transitions matrices elements of each column in a vector (i.e., a_{1i}, \dots, a_{Ni} are stored in the vector $Transition_i$). Also, to store the α_t , in such a way to store elements of the same column in separate vector (i.e., $\alpha_i(1), \dots, \alpha_i(N)$ is stored in the vector $Alpha_i$).

2) The use of MapReduce paradigm [11] for partitioning the sequence into blocks. It enables parallel distributed processing of large sets of data, converting them into another set of data (map function) and then combining and reducing those output sets of data into smaller sets of data (reduce function). It allows to apply RDDs transformations including several MapReduce-like operations (e.g., map, reduce, collect).

3) The use of broadcast variables to increase the performance and reduce the communication costs. Spark attempts to effectively distribute broadcast variables using powerful broadcast algorithms [41]. They allow to keep a read-only variable cached on each machine rather than shipping a copy of it with tasks. Thus, broadcast makes it possible to distribute vectors or matrices of parameters on all nodes. In our case, transition matrix, emission probabilities and initial probabilities are broadcasted (see Fig. 2b).

Now, we describe, step by step, the implementation of ParaDist-Forward algorithm:

(1) Initialization step: each executor will execute an initialization task of a $\alpha_1(j)$ for a given j , $1 \leq j \leq N$.

for each *executor*_{*j*} of *N* executors do

$$\alpha_1(j) \leftarrow \pi_j b_j(o_1)$$

end for

This operation is described in Fig. 2d.

So, the initialization step has a complexity of $O(1)$ instead of $O(N)$.

For HMM with multiple observations (M), we will have to use $N * M$ executors in parallel.

(2) Induction step: at each time t , for the calculation of $\alpha_{t+1}(j)$, we must first calculate the $\alpha_t(j)$. So, since $\alpha_t(j)$ depend on time, we cannot parallelize over t , but it is possible over N (states number).

for $t \leftarrow 1$ to $T - 1$ do

for each *executor*_{*j*} of *N* executors do

for each *executor*_{*i*} of *N* executors do

$$\alpha_{t+1}(j) \leftarrow b_j(o_{t+1}) \sum_{i=1}^N \alpha_t(i) a_{ij}$$

end for

end for

end for

The calculation process can be schematized as in Fig. 2c.

(3) Termination step: now, we have all $\alpha_T(i)$ stored in the vector $Alpha_T$, we can simply use Spark's RDD action 'reduce' to sum all elements of the vector (Fig. 2e).

$$Pr\{O|\lambda\} \leftarrow Alpha_T.reduce(lambda a, b : a + b)$$

The proposed parallel distributed forward algorithm using Spark (ParaDist-Forward) is presented in Algorithm 1. In backward algorithm, we use the same principle as forward variable. ParaDist-Backward algorithm is presented in Algorithm 2. Baum-Welch algorithm has a complexity of $O((T - 1)N^2)$, with the proposed implementation, we were able to reduce this complexity to $O(T - 1)$. The proposed parallel distributed Baum-Welch algorithm using Spark (ParaDist-Baum-Welch) is presented in Algorithm 3.

Algorithm 1: ParaDist-Forward Algorithm

input : A model $\lambda = (A, B, \Pi)$, a sequence of observations $O = o_1, o_2, \dots, o_T$

output: The probability $Pr\{O | \lambda\}$

1 **begin**

2 **for each** *executor*_{*j*} of *N* executors **do**

3 **parallel do**

4 $\alpha_1(j) \leftarrow \pi_j b_j(o_1) \{j \in \{1, 2, 3, \dots, N\}\}$

5 **for** $t \leftarrow 1$ to $T - 1$ **do**

6 **for each** *executor*_{*i,j*} of $N*N$ executors **do**

7 **parallel do**

8 calculate(*map*) $\alpha_t(i) a_{ij}$ and store $\alpha_t(i)$ in $Alpha_t \{i, j \in \{1, 2, 3, \dots, N\}\}$

9 make the sum (*reduce*) $\alpha_t(i) a_{ij}$, then multiply by

$b_j(o_{t+1}) \{i, j \in \{1, 2, 3, \dots, N\}\}$

10 $Pr\{O | \lambda\} \leftarrow Alpha_T.reduce(lambda a, b : a + b)$

11 **return** $Pr\{O | \lambda\}$

B. Proposed Architecture for Modeling and Solving Big Data Analytics Problems using ParaDist-HMM Model and Spark

The proposed approach, described in Fig. 3, is based on use of Apache Spark offering Spark core for batch processing, Spark streaming for real time processing and Spark sql for connection to other applications and data exploration. Next, in this section, we will present the main steps of the proposed Spark-based architecture for modeling and analyzing big data using ParaDist-HMM.

Spark is an open source big data processing framework built to perform advanced analysis. It has several advantages over other big data technologies like Hadoop and Storm. Spark offers a complete and unified framework to meet the needs of big data processing and analysis for various datasets (see Fig. 4a). It allows applications on Hadoop clusters to be executed up to 100 times faster in memory and 10 times faster on disk. Spark is composed of seven elements: Spark core of

Algorithm 2: ParaDist-Backward Algorithm

input : A model $\lambda = (A, B, \Pi)$, a sequence of observations $O = o_1, o_2, \dots, o_T$
output: The probability $Pr\{O | \lambda\}$

```

1 begin
2   for each executorj of N executors do
3     parallel do
4        $\beta_T(j) \leftarrow 1 \{j \in \{1, 2, 3, \dots, N\}\}$ 
5     for  $t \leftarrow T - 1$  downto 1 do
6       for each executori,j of N*N executors do
7         parallel do
8           calculate  $\beta_{t+1}(j)a_{ij}b_j(o_{t+1})$  and store  $\beta_t(j)$  in  $Beta_t \{i, j \in \{1, 2, 3, \dots, N\}\}$ 
9       for each executorj of N executors do
10        parallel do
11          calculate  $\pi_i b_i(o_1)\beta_1(i) \{i \in \{1, 2, 3, \dots, N\}\}$ 
12         $Pr\{O | \lambda\} \leftarrow \text{sum}(\pi_i b_i(o_1)\beta_1(i))$ 
13      return  $Pr\{O | \lambda\}$ 

```

data engine, Spark cluster manager (includes Hadoop, Apache Mesos and built-in Standalone cluster manger), Spark SQL, Spark streaming, Spark machine learning library MLlib, Spark GraphX and Spark programming tools.

The steps of the Spark-based architecture for modeling and analyzing big data using ParaDist-HMM are the following:

Step 1: Data collection and data storage

For data ingestion, we used Sqoop (Fig. 4b) to import structured data from HBase, Hive or Hadoop Distributed File System (HDFS). For data streaming, we used Kafka (Fig. 4c) to collect the data streaming. It works in combination with Spark for real-time analysis and rendering of streaming data used. Data are, then, loaded in HDFS (Fig. 4d).

For cluster management, we used Spark on Hadoop YARN cluster (Fig. 4e). This coordinates data ingestion from Sqoop and Kafka and other services that deliver data into Spark cluster. YARN cluster manager (Fig. 4f) allows dynamic sharing and central configuration of the same pool of cluster resources between various frameworks that run on YARN. The number of executors to use can be selected by the user unlike the Standalone mode. When executing a program on top of Spark, it runs as a driver. The driver passes execution of parallel operations such as map or reduce to Spark.

Step 2: Feature selection and extraction

The *mllib.feature* package contains several classes for common feature transformations. These include algorithms to construct feature vectors from text (or other tokens) and ways to normalize and scale features.

STEP 3: Machine learning algorithms

In this step, we go through the learning machine algorithms to solve big data analytics problems thanks to the Spark's machine learning library, MLlib in addition to the proposed implementation under Spark of HMMs, ParaDist-HMM.

STEP 4: Model evaluation

When building machine learning models, we need to evaluate the performance of the model on some criteria. *spark.mllib* provides a suite of metrics for the purpose of evaluating the performance of machine learning models.

Algorithm 3: ParaDist-Baum-Welch algorithm

input : Initial model $\lambda = (A, B, \Pi)$, a sequence of observations O
output: Optimal Model parameters
 $\bar{A} = \{a_{ij}\}, \bar{B} = \{b_j(v_k)\}, \bar{\Pi} = \{\pi_i\}$

```

1 Begin
2   for each executorj of N executors do
3     parallel do
4        $\alpha_1(j) \leftarrow \pi_j b_j(o_1) \{j \in \{1, 2, 3, \dots, N\}\}$ 
5   for  $t \leftarrow 1$  to  $T - 1$  do
6     for each executori,j of N*N executors do
7       parallel do
8         calculate  $(map) \alpha_t(i)a_{ij}$  and store  $\alpha_t(i)$  in  $Alpha_t \{i, j \in \{1, 2, 3, \dots, N\}\}$ 
9         sum (reduce)  $\alpha_t(i)a_{ij}$ , then multiply by  $b_j(o_{t+1})\{i, j \in \{1, 2, 3, \dots, N\}\}$ 
10       $Pr\{O | \lambda\} \leftarrow Alpha_T \cdot \text{reduce}(\text{lambda } a, b : a + b)$ 
11     for each executorj of N executors do
12       parallel do
13          $\beta_T(j) \leftarrow 1 \{j \in \{1, 2, 3, \dots, N\}\}$ 
14     for  $t \leftarrow T - 1$  downto 1 do
15       for each executori,j of N*N executors do
16         parallel do
17           calculate  $\beta_{t+1}(j)a_{ij}b_j(o_{t+1})$  and store  $\beta_t(j)$  in  $Beta_t \{i, j \in \{1, 2, 3, \dots, N\}\}$ 
18     for each executort,i of T*N executors do
19       parallel do
20         calculate  $\gamma_t(i) \leftarrow \alpha_t(i)\beta_t(i)/Pr\{O | \lambda\}$  and store  $\gamma_t(i)$  in  $Gamma_t \{i \in \{1, 2, 3, \dots, N\}; t \in \{1, 2, 3, \dots, T\}\}$ 
21     for each executort,i,j of (T-1)*N*N executors do
22       parallel do
23         calculate  $\xi_t(i, j) \leftarrow \alpha_t(i)a_{ij}\beta_{t+1}(j)b_j(o_{t+1})/Pr\{O | \lambda\}$  and store  $\xi_t(i, j)$  in  $X_{it} \{i, j \in \{1, 2, 3, \dots, N\}; t \in \{1, 2, 3, \dots, T - 1\}\}$ 
24     for each executori,j of N*N executors do
25       parallel do
26         calculate  $\bar{a}_{ij} \leftarrow \text{sum}(\xi_t(i, j))/\text{sum}(\gamma_t(i)) \{i, j \in \{1, 2, 3, \dots, N\}; t \in \{1, 2, 3, \dots, T - 1\}\}$ 
27     for each executorj,k of N*M executors do
28       parallel do
29         calculate  $\bar{b}_j(v_k) \leftarrow \text{sum}(\gamma_t(j))/\text{sum}(\gamma_t(j)) \{o_t = v_k; j \in \{1, 2, 3, \dots, N\}; k \in \{1, 2, 3, \dots, M\}; t \in \{1, 2, 3, \dots, T\}\}$ 
30     for each executori of N executors do
31       parallel do
32         calculate  $\bar{\pi}_i \leftarrow \gamma_1(i) \{i \in \{1, 2, 3, \dots, N\}\}$ 
33     set  $\lambda \leftarrow \bar{\lambda}$  and go to 18 unless some convergence criterion is met
34     return  $\bar{A} = \{a_{ij}\}, \bar{B} = \{b_j(v_k)\}, \bar{\Pi} = \{\pi_i\}$ 

```

V. MATERIAL AND METHODS

In this section, we give a description of the dataset used and we present the experimental setup and the architectural configuration of the experiments.

A. Experiments Data

We performed various experiments to solve fundamental problems of HMM based on datasets that we have selected to be representative of the main field of application of HMM. In the experiments, we firstly, used synthetic data, since they allow better understanding of the real data and identifying the special features of it for a considerable number of use cases. They also help, by simulating real data sets, to fulfill their gaps. Generating synthetic data also helps to get a view of how a larger dataset would be, this view could save us from getting a very large dataset and avoid a lot of work effort that may require. In addition, synthetic data allow to know if a model would be useful with the data by providing early results with the synthetic data, giving a performance preview without needing to retrieve more real data [42]. The synthetic data were generated using PyMC3 HMM [43], an open-source probabilistic programming package written in Python, giving the parameters of the model consisting of sequences of integers drawn from a multinomial distribution. We assume to have an ergodic HMM. First, we choose the initial HMM parameters randomly in such a way the initial state probabilities, the state transition probabilities and the symbol probabilities satisfy the following criteria: $\sum_{i=1}^N \pi_i = 1$, $\sum_{j=1}^N a_{ij} = 1$ and $\sum_{k=1}^M b_j(v_k) = 1$. An adequate choice for Π , A and B is to assign to each state transition probability a_{ij} a real value at random between 0 and $1/N$, a set of random values between 0 and $1/N$ to each initial state probability π_i and a random value between 0 and $1/M$ to each symbol probability $b_j(v_k)$. Then, as reported in [44], given appropriate values of N , M , A , B and Π , the HMM is used to generate an observation sequence $O = o_1, \dots, o_T$ as follows: 1- Choose an initial state $q_t = S_i$ according to the initial state distribution π_i . 2- Set $t = 1$. 3- Choose $O_t = v_k$ according to the symbol probability distribution in state S_i , i.e., $b_i(v_k)$. 4- Transit to a new state $q_{t+1} = S_j$ according to the transition probability distribution for state S_i , i.e., a_{ij} . 5- Set $t = t + 1$; return to step 3- if $t < T$; otherwise terminate the procedure.

Then, in order to evaluate the prediction accuracy of algorithms, we used a real financial dataset consisting of daily data from the Dow Jones Industrial Average (DJIA) stock market index during the period between January 1, 2010 and July 1, 2020 obtained from Yahoo Finance website [45].

B. Experimental Setup

For experimental evaluations, we have chosen scenarios that reflect as much as possible a real world of big data analytic.

In the first scenario, experiments are conducted in Amazon EC2 Elastic Compute Cloud using *t2.large* cloud computing platform with 8 GB of memory and 2 CPU with 2.0.1 as version of Spark with 5 GB of storage for Amazon S3. In the second scenario, we perform the evaluation in a pseudo-distributed mode with 3 local machines (a laptop Acer aspire

5551g-p324g32mnkk with an AMD athlon II dual core processor p320, 2.3 GHz, 4Go ddr4 and on an integrated ati radeon hd5470 512Mo graphics card based on the park xt graphics processor, a laptop HP 620 with an intel core2 duo processor T6570, 2,10 GHz, a memory 4GB ddr3 1333MHz sdram, 320 GB hdd and a graphics card mobile intel gma 4500mhd and an Acer extensa tower pc workstation em2610 i5-4460 4th gen intel core i5 4 GB ddr3-sdram 500 GB hdd freedOS PC black). In the third scenario, we used Spark in single node (laptop Acer aspire 5551g-p324g32mnkk) mode so, we can implement the classic algorithms of HMMs. The experiments reported in this paper were performed on Ubuntu Linux 18.04.5 LTS with the Linux Kernel 5.4. For BW algorithm, in each experiment, we randomly selected from the database a training dataset consisting of 80 % of data and a tests dataset representing a percentage of 20%. Several experiments were performed independently.

VI. RESULTS AND DISCUSSION

In this paper, the results obtained after performing different experiments are evaluated, based on the comparison between the classical algorithm and the proposed algorithm (i.e., ParaDist-HMM) in a pseudo distributed environment and in a cloud environment, in terms of running time, speedup and accuracy using synthetic and real. In this section we give a detailed description of different experimental evaluations performed in this study followed by an analysis of the results.

A. Running Time

To investigate and examine the total running time, we have conducted several experiments varying the number of states and the number of sequences. Each experiment was repeated 3 times and the running time is the average of the running times of the three tests. We show the running time in terms of data sizes (i.e., sequences number) and states numbers. We compute the running time for different values of states numbers (i.e., 10, 100, 1000, 5000, 7000 and 10000). Fig. 5a illustrates the running time taken for the ParaDist-Forward algorithm as it varies with the states number. We can see a clear improvement since the time complexity is optimized. In the second experiment, we compute the running time varying the number of sequences with values ranging from 10 up to 5000000. Fig. 5b shows ParaDist-Forward algorithm performance in terms of running time according to sequences number. Concerning BW algorithm, Fig. 6a shows ParaDist-Baum-Welch algorithm performance in terms of running time according to states number. While Fig. 6b shows how the running time of BW algorithm varies with the number of sequences. From the curves on these figures, we can see a significant improvement in the running time in terms of states number and sequences number, the difference is very clear between the parallel distributed Baum-Welch and the conventional one. We notice that increasing the states number and sequences number (dataset size) has a positive effect on the amelioration of running time.

B. Speedup

Speedup is one of the main parallel performance metrics which measures the evolution of the execution time according to the number of nodes. It measures acceleration, the benefit

obtained by an algorithm and a parallel implementation compared to the same algorithm on a single node. Fig. 7a presents a comparison between ParaDist-Forward algorithm in cloud environment with 20 nodes, ParaDist-Forward algorithm in pseudo distributed environment with 5 nodes and the classical algorithm implemented in a single node on a local machine. This figure clearly shows an excellent performance of the proposed algorithm especially in a cloud environment. It is also noted that as long as the number of states or the number of sequences becomes important, the result is better. From these results, we can see that the proposed algorithm is positively affected by the size of the input data and the number of nodes, hence its high scalability.

We also performed the classical implementation of the Baum-Welch algorithm in a single node and the proposed algorithm, ParaDist-Baum-Welch, in a pseudo distributed (5 nodes) and a distributed environment (20 nodes). The comparison results are shown in the Fig. 7b. We can deduce that the proposed version of Baum-Welch algorithm in a parallel distributed environment presents a great improvement if we compare it with the results of the implementation of the classical Baum-Welch algorithm in a single node. For a more meaningful evaluation, we compared ParaDist-Forward and ParaDist-Baum-Welch algorithms to those implemented under Mahout MapReduce using the package *org.apache.mahout.classifier.sequencelearning.hmm* as function of data size and nodes number. From Table II showing the speedup percent comparison, we observe the superiority of the improved algorithm compared to the classical version, where both implementations (i.e., ParaDist-Forward and MapReduce's) are affected by the increase of data size since we observe a decrease in the speedup. For small data sizes, the proposed algorithm outperforms that of MapReduce by up to three times and a half. However, as compared to MapReduce's, the proposed algorithm surpasses it up to two and a half times for large data sizes. The Table III shows the results of the acceleration comparison of both versions according to the number of sequences and the number of nodes. This table shows a clear improvement in terms of running time. We can also notice that this increase is proportional to the number of sequences and to the number of nodes. For small data sizes the speedup is not too high and we also observe that for this data case the classic algorithm outperforms even the proposed algorithm for a low number of nodes. The ratio between the speedup of the proposed algorithm and that of MapReduce is of the order of 10. For large data sizes, our algorithm surpasses that of MapReduce up to two times.

Finally, we also compared ParaDist-Forward to the main proposed models in the literature in terms of speedup. Due to the problem difference, the model parameters for different run in this comparison might be different, thus we did not directly compute the running time of each algorithm. Since both the serial forward and the proposed parallel version in each paper were executed using the same dataset with the same parameters, we compute the relative speedup between the two in each case and compare it over the other versions. Table IV shows the result of average relative speedup comparison of ParaDist-Forward algorithm compared to those of [32], [33], [34], [35], [36], [37] and [38]. The results show that the speedup of the proposed model has the best results compare to the benchmark models.

TABLE II. FORWARD ALGORITHM SPEEDUP %

Sequences number	5 nodes		20 nodes	
	Parad Dist-Forward	Map Re-duce's	Parad Dist-Forward	Map Re-duce's
7000	11,50	3,28	2880,02	1152,01
1000000	5,26	2,10	96,15	38,46
2000000	5,55	2,23	111,11	44,45
3000000	5,95	2,41	120,05	48,33

TABLE III. BAUM-WELCH ALGORITHM SPEEDUP %

Sequences number	5 nodes		20 nodes	
	Para Dist-Baum-Welch	Map Re-duce's	Para Dist-Baum-Welch	Map Re-duce's
1000	0,31	3,08	3,01	30,15
500000	5,01	10,06	29,65	59,31
800000	5,50	10,87	255,35	512,63
1000000	6,01	11,95	260,86	518,57

TABLE IV. SPEEDUP FACTOR COMPARISON OF FORWARD ALGORITHM

	ours	[32]	[33]	[34]	[35]	[36]	[37]	[38]
Average Speedup Factor	5333.34x	9.2x	180x	3x	4x	880x	3.5x	1.1x

C. Accuracy

As we mentioned above, the data are divided into two groups: a training dataset consisting of 80 % of data and a tests dataset representing a percentage of 20%. Our primary goal is to investigate how the prediction accuracy of the HMMs learned using different versions of Baum-Welch algorithm varies as function of the number of iterations, in terms of data size and as function of the number of nodes. We compared the quality of prediction of the HMMs with Baum-Welch algorithm in the conventional and the parallel distributed versions, using the occurrences of the output values correctly predicted. To assess the HMM performance, we used two metrics: the accuracy and the Root Mean Square Error (RMSE). The accuracy is defined as the number of correctly predicted values under the total number values in the testing set. The RMSE of a model prediction measures the difference between the values predicted by a model and the values actually observed. The RMSE is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_i^n (V_{observed} - V_{predicted})^2}{n}}$$

where $V_{observed}$ is the observed value and $V_{predicted}$ is the predicted value at time i and n is the total number of test data. Table V shows the prediction accuracy for HMM learned by the conventional BW on a single node and HMM learned by the ParaDist-Baum-Welch in a distributed environment. This table illustrates how the prediction accuracy of the models varies for different values of iterations numbers. We note, here,

TABLE V. HMM ACCURACY (%) VS ITERATIONS NUMBER

Number of Iterations	ParaDist-HMM	Conventional HMM
100	87.01	87.01
600	87.08	87.08
1000	91.18	91.05
10000	93.56	93.50
100000	96.67	94.34

that we used, for the prediction accuracy evaluation, the financial dataset from the DJIA index in order to forecast financial market behavior. We observe an improvement in the prediction accuracy with the increase in the number of iterations. We also investigated how well the learning algorithm affect the prediction accuracy of the model as function of the number of sequences.

Table VI shows the change in RMSE of HMM model prediction with different data size for HMM learned by ParaDist-Baum-Welch and conventional BW. As the number of sequences increases, a slight decrease in the accuracy of the models appears in both scenarios. But the difference in RMSE values for high numbers of sequences indicates a difference on how accurately the models predict the output. The HMM trained using ParaDist-Baum-Welch clearly outperforms the other model. Like shown in Table V, our model achieve comparable accuracy to the classical one for lower numbers of iterations and presents a best prediction accuracy of 96.67% for a number of iterations equal to 100000, while the RMSE of the model prediction achieves 3,850 as shown in Table VI. The results indicate the proposed model is more accurate and provide good estimation for large numbers of iterations for big data sizes since the increase in the number of iterations, the refinement of the model improves and therefore the learning phase which explains the good results of the model.

We, finally, also compared our model to the main proposed models in the literature. Table VII presents the results of prediction quality comparison of our ParaDist-Baum-Welch algorithm compared to those of Mahout MapReduce, [31] and [32]. In this table, we compare the average prediction accuracy achieved by this algorithm in an identical scenario. As we can see, our algorithm gives almost the same result as that of MapReduce and outperforms other benchmark algorithms in terms of prediction accuracy. Although there is a minor fluctuation in the accuracy for a lower number of iterations or for small data, this is due to the random nature of the choice of initial parameters and model topology and does not affect the analysis to a large extent. A subject around the HMMs which certainly remains interesting to explore. Nonetheless, the ParaDist-HMM meets minimum benchmarks for accuracy, often outperforming the conventional HMM mainly in a big data context.

VII. CONCLUSION AND OUTLOOK

In this paper, we presented ParaDist-HMM model which consists of new parallel distributed versions for main HMM algorithms. To put this implementation into practice, we have proposed a Spark-based architecture for big data analytics by fully exploiting the benefits of this framework with a set

TABLE VI. HMM PREDICTION RMSE VS SEQUENCES NUMBER

Number of Sequences	ParaDist-HMM	Conventional HMM
1200	2.115	2.213
6000	3.370	2.972
80000	3.566	3.215
100000	3.825	3.256
1200000	3.850	3.331

TABLE VII. PREDICTION QUALITY COMPARISON (%)

	ours	MapReduce's	[31]	[32]
Average Prediction Accuracy	96.43	96.41	92.04	92

of powerful tools for managing and analyzing big data. In summary, the results of the various experiments carried out on synthetic data and real financial data show that the proposed parallel distributed algorithms using Spark outperforms the classics and the other main solutions presented previously in the literature in terms of running time and speedup. As for Baum-Welch algorithm, our approach, indeed, improves the learning accuracy leading to better learning performance. The proposed ParaDist-HMM model is well suited to the big data analytics problems, since it has shown good performance for a very large amount of data and have proven to be robust and efficient in terms of processing speed, execution time, accuracy and scalability.

As a continuation of this work, we will deal with the decoding problem for the HMMs. It is also necessary to study continuous-time HMM case by focusing on the fundamental problem of HMM which is the training problem. It would also be important to address the case of multiple observations. Naturally, it would be interesting to apply our results to other time series problems mainly for modeling and forecasting other financial time series, bioinformatics and medicine problems and natural language processing problems.

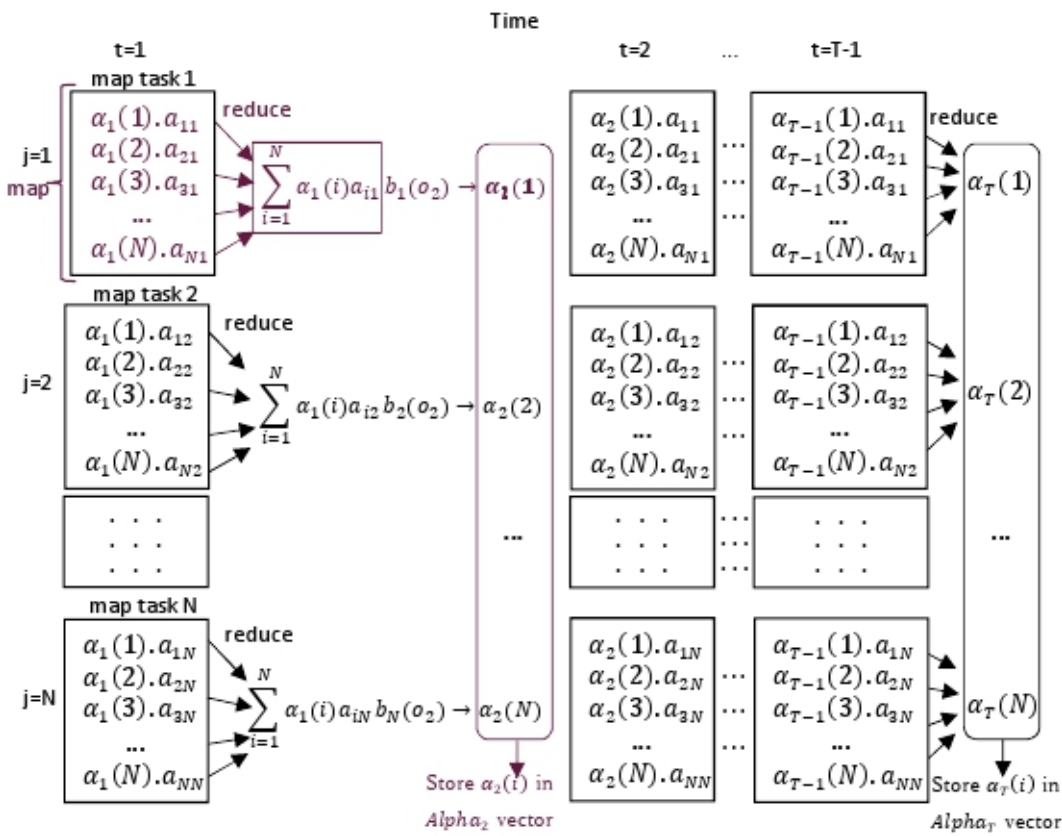
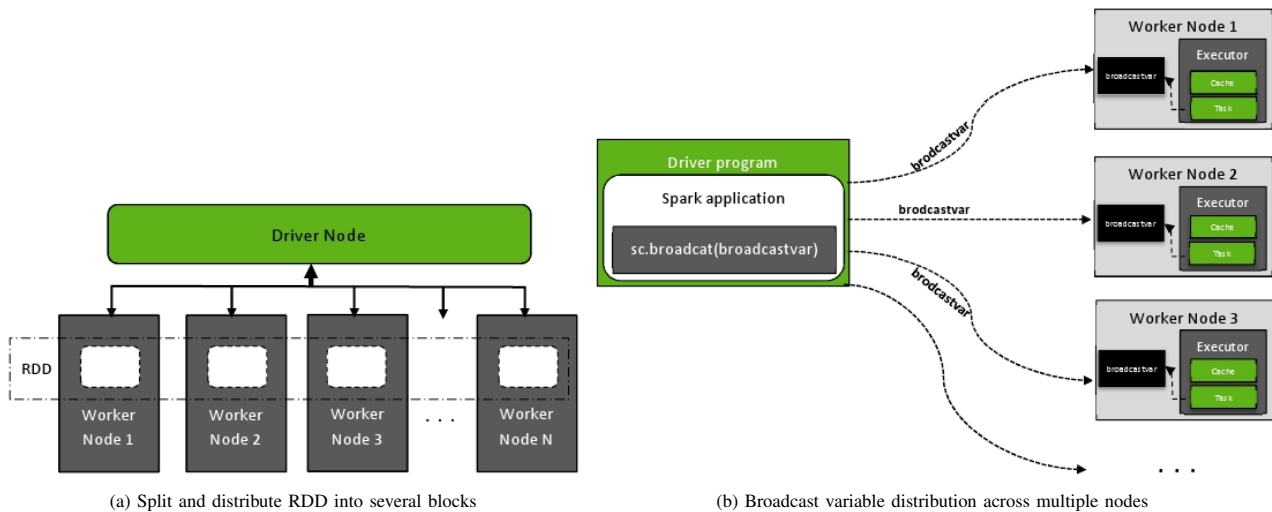
As future work, some promising directions include studying possible combinations between hidden Markov models and fuzzy models or some deep learning algorithms or metaheuristics techniques or to use cascading methods to improve the obtained results. Future work will also focus on using other metrics to properly evaluate these algorithms.

REFERENCES

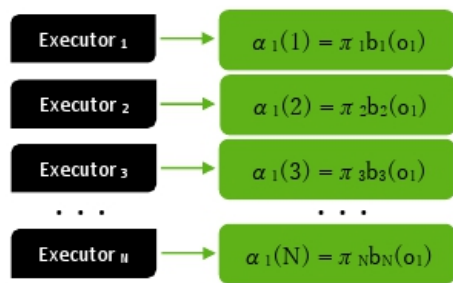
- [1] A. H. Hussein, *Internet of things (IOT): Research challenges and future applications*, International Journal of Advanced Computer Science and Applications, vol. 10, no 6, p. 77-82, 2019.
- [2] I. Sassi, S. Anter and A. Bekkhoucha, *An Overview of Big Data and Machine Learning Paradigms*, In : International Conference on Advanced Intelligent Systems for Sustainable Development. Springer, Cham, p. 237-251, 2018.
- [3] C. C. Qi, *Big data management in the mining industry*, International Journal of Minerals, Metallurgy and Materials, vol. 27, no 2, p. 131-139, 2020.
- [4] G. T. Reddy, M. P. Reddy, K. Lakshmana, et al., *Analysis*

- of dimensionality reduction techniques on big data, IEEE Access, vol. 8, p. 54776-54788, 2020.
- [5] I. Sassi, S. Ouafitouh and S. Anter, *Adaptation of Classical Machine Learning Algorithms to Big Data Context: Problems and Challenges: Case Study: Hidden Markov Models Under Spark*, In : 2019 1st International Conference on Smart Systems and Data Science (ICSSD). IEEE, p. 1-7, 2019.
- [6] D. P. Acharjya and K. Ahmed, *A survey on big data analytics: challenges, open research issues and tools*, International Journal of Advanced Computer Science and Applications, vol. 7, no 2, p. 511-518, 2016.
- [7] M. A. Hashmani, S. M. Jameel, A. M. Ibrahim, M. Zaffar and K. Raza, *An ensemble approach to big data security (cyber security)*, International Journal of Advanced Computer Science and Applications, vol. 9, no 9, p. 75-77, 2018.
- [8] V. Belov, A. Tatarintsev and E. Nikulchev, *Choosing a Data Storage Format in the Apache Hadoop System Based on Experimental Evaluation Using Apache Spark*, Symmetry, vol. 13, no 2, p. 195, 2021.
- [9] A. Ashabi, S. B. Sahibuddin and M. S. Haghghi, *Big Data: Current Challenges and Future Scope*, In : 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE). IEEE, p. 131-134, 2020.
- [10] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García and F. Herrera, *Big data preprocessing*, Cham: Springer, 2020.
- [11] I. Sassi and S. Anter, *A study on big data frameworks and machine learning tool kits*, In Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019, pp. 61-68, 2019.
- [12] A. Mostafaeipour, A. Jahangard Rafsanjani, M. Ahmadi and J. Arockia Dhanraj, *Investigating the performance of Hadoop and Spark platforms on machine learning algorithms*, The Journal of Supercomputing, p. 1-28, 2020.
- [13] F. Ameer, M. K. Hanif, R. Talib, M. U. Sarwar, Z. Khan, K. Zulfiqar and A. Riasat, *Techniques, Tools and Applications of Graph Analytic*, International Journal of Advanced Computer Science and Applications, vol. 10, no 4, p. 354-363, 2019.
- [14] A. K. Gupta, P. Varshney, A. Kumar, B. R. Prasad and S. Agarwal, *Evaluation of mapreduce-based distributed parallel machine learning algorithms*, In : Advances in Big Data and Cloud Computing. Springer, Singapore, p. 101-111, 2018.
- [15] T. Y. Liu, W. Chen and T. Wang, *Distributed machine learning: Foundations, trends, and practices*, In : Proceedings of the 26th International Conference on World Wide Web Companion, p. 913-915, 2017.
- [16] Z. H. Zhan, J. Zhang, Y. Lin, J. Y. Li, T. Huang, X. Q. Guo, F. Wei, S. Kwong, X. Zhang and R. You, *Matrix-Based Evolutionary Computation*, IEEE Transactions on Emerging Topics in Computational Intelligence, 2021.
- [17] M. A. Amin, M. K. Hanif, M. U. Sarwar, A. Rehman, F. Waheed and H. Rehman, *Parallel Backpropagation Neural Network Training Techniques using Graphics Processing Unit*, International Journal of Advanced Computer Science and Applications, vol. 10, no 2, p. 563-566, 2019.
- [18] D. R. Westhead and M. S. Vijayabaskar, *Hidden Markov Models*, Springer Science+ Business Media LLC, 2017.
- [19] G. S. Grimmett, *Probability and random processes*, Oxford university press, 2020.
- [20] W. Zucchini, I. L. MacDonald, R. Langrock, *Hidden Markov models for time series: an introduction using R*, CRC press, 2017.
- [21] L. Rabiner and B. Juang, *An introduction to hidden Markov models*, IEEE ASSP Magazine, vol. 3, no 1, p. 4-16, 1986.
- [22] J. Lember and J. Sova, *Existence of infinite Viterbi path for pairwise Markov models*, Stochastic Processes and their Applications, vol. 130, no 3, p. 1388-1425, 2020.
- [23] K. Slavakis, G. B. Giannakis and G. Mateos, *Modeling and optimization for big data analytics: (statistical) learning tools for our era of data deluge*, IEEE Signal Processing Magazine, vol. 31, no 5, p. 18-31, 2014.
- [24] S. Alshamrani, Q. Waseem, A. Alharbi, W. Alosaimi, H. Turabieh and H. Alyami, *An Efficient Approach for Storage of Big Data Streams in Distributed Stream Processing Systems*, International Journal of Advanced Computer Science and Applications, vol. 11, no 5, p. 91-98, 2020.
- [25] H. Abounaser, I. Talkhan and A. Fahmy, *A Parallel Fuzzy-Genetic Algorithm for Classification and Prediction*, International Journal Of Advanced Computer Science and Applications, vol. 7, no 10, p. 161-171, 2016.
- [26] J. Verbaeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen and J. S. Rellermeier, *A survey on distributed machine learning*, ACM Computing Surveys (CSUR), vol. 53, no 2, p. 1-33, 2020.
- [27] L. Zhou, S. Pan, J. Wang and A. V. Vasilakos, *Machine learning on big data: Opportunities and challenges*, Neurocomputing, vol. 237, p. 350-361, 2017.
- [28] V. K. Gunjan, J. M. Zurada, B. Raman and G. R. Gangadharan, *Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough*, Springer International Publishing, 2020.
- [29] M. Bhattacharya, *Expensive optimisation: A metaheuristic perspective*, International Journal Of Advanced Computer Science and Applications, vol. 4, no 1, p. 203-209, 2013.
- [30] S. Russell and P. Norvig, *Artificial intelligence: a modern approach*, 2002.
- [31] X. Ma, D. Schonfeld and A. Khokhar, *Distributed multi-dimensional hidden Markov model: theory and application in multiple-object trajectory classification and recognition*, In : Multimedia Content Access: Algorithms and Systems II. International Society for Optics and Photonics, p. 682000, 2008.
- [32] L. Yu, Y. Ukidave and D. Kaeli, *GPU-accelerated HMM for Speech Recognition*, In : 2014 43rd International Conference on Parallel Processing Workshops. IEEE, p. 395-402, 2014.
- [33] S. Hymel, *Massively parallel hidden Markov models for wireless applications*, Doctoral dissertation. Virginia Tech, 2011.
- [34] A. Sand, C. N. Pedersen, T. Mailund and A. T. Brask, *HMMLib: A C++ library for general hidden Markov models exploiting modern CPUs*, In : 2010 Ninth International Workshop on Parallel and Distributed Methods in Verification, and Second International Workshop on High Performance Computational Systems Biology. IEEE, p. 126-134, 2010.
- [35] J. Nielsen and A. Sand, *Algorithms for a parallel im-*

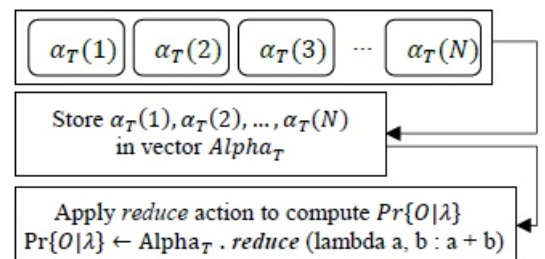
- plementation of hidden Markov models with a small state space, In : 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum. IEEE, p. 452-459, 2011.
- [36] C. Liu, *cuHMM: a CUDA implementation of hidden Markov model training and classification*, The Chronicle of Higher Education, p. 1-13, 2009.
- [37] J. Li, S. Chen and Y. Li, *The fast evaluation of hidden Markov models on GPU*, In : 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. IEEE, p. 426-430, 2009.
- [38] C D. Mitchell, L. H. Jamieson, M. P. Harper and R. Helzerman, *Implementing a hidden Markov model with duration modeling on the MasPar MP-1*, ECE Technical Reports, p. 190, 1994.
- [39] R. Bosagh Zadeh, X. Meng, A. Ulanov, B. Yavuz, L. Pu, S. Venkataraman, E. Sparks, A. Staple and M. Zaharia, *Matrix computations and optimization in apache spark*, In : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 31-38, 2016.
- [40] R. Gu, Y. Tang, Z. Wang, S. Wang, X. Yin, C. Yuan and Y. Huang, *Efficient large scale distributed matrix computation with spark*, In : 2015 IEEE International Conference on Big Data (Big Data). IEEE, p. 2327-2336, 2015.
- [41] M. Armbrust, T. Das, A. Davidson, A. Ghodsi, A. Or, J. Rosen, I. Stoica, P. Wendell, R. Xin and M. Zaharia, *Scaling spark in the real world: performance and usability*, Proceedings of the VLDB Endowment, vol. 8, no 12, p. 1840-1843, 2015.
- [42] J. Ferrando Huertas, *Generating synthetic data through Hidden Markov Models*, 2018.
- [43] J. Salvatier, T. V. Wiecki and C. Fonnesbeck, *Probabilistic programming in Python using PyMC3*, PeerJ Computer Science, vol. 2, p. e55, 2016.
- [44] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, vol. 77, no 2, p. 257-286, 1989.
- [45] Yahoo!, *Dow Jones Industrial Average*, finance.yahoo.com. <https://finance.yahoo.com/quote/%5edji/> (accessed Feb. 1, 2020).



(c) Parallel distributed calculation of forward variable



(d) Paralyzied initialization on Spark



(e) Computation of the probability $\text{Pr}\{O|\lambda\}$

Fig. 2. Implementation Steps of ParaDist-Forward Algorithm.

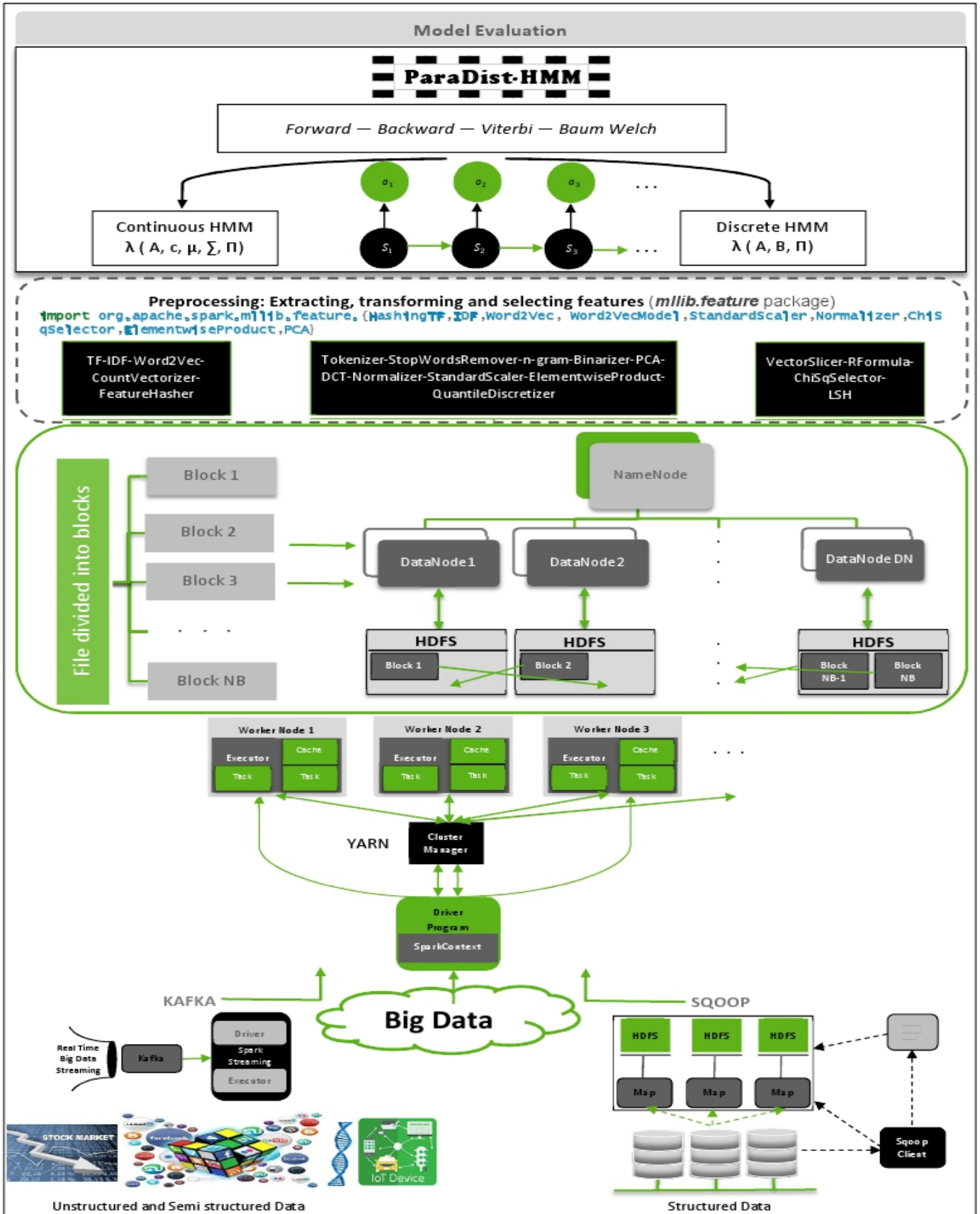


Fig. 3. Overview of Proposed Approach for Modeling and Solving Big Data Analytics Problems using ParaDist-HMM and Spark.

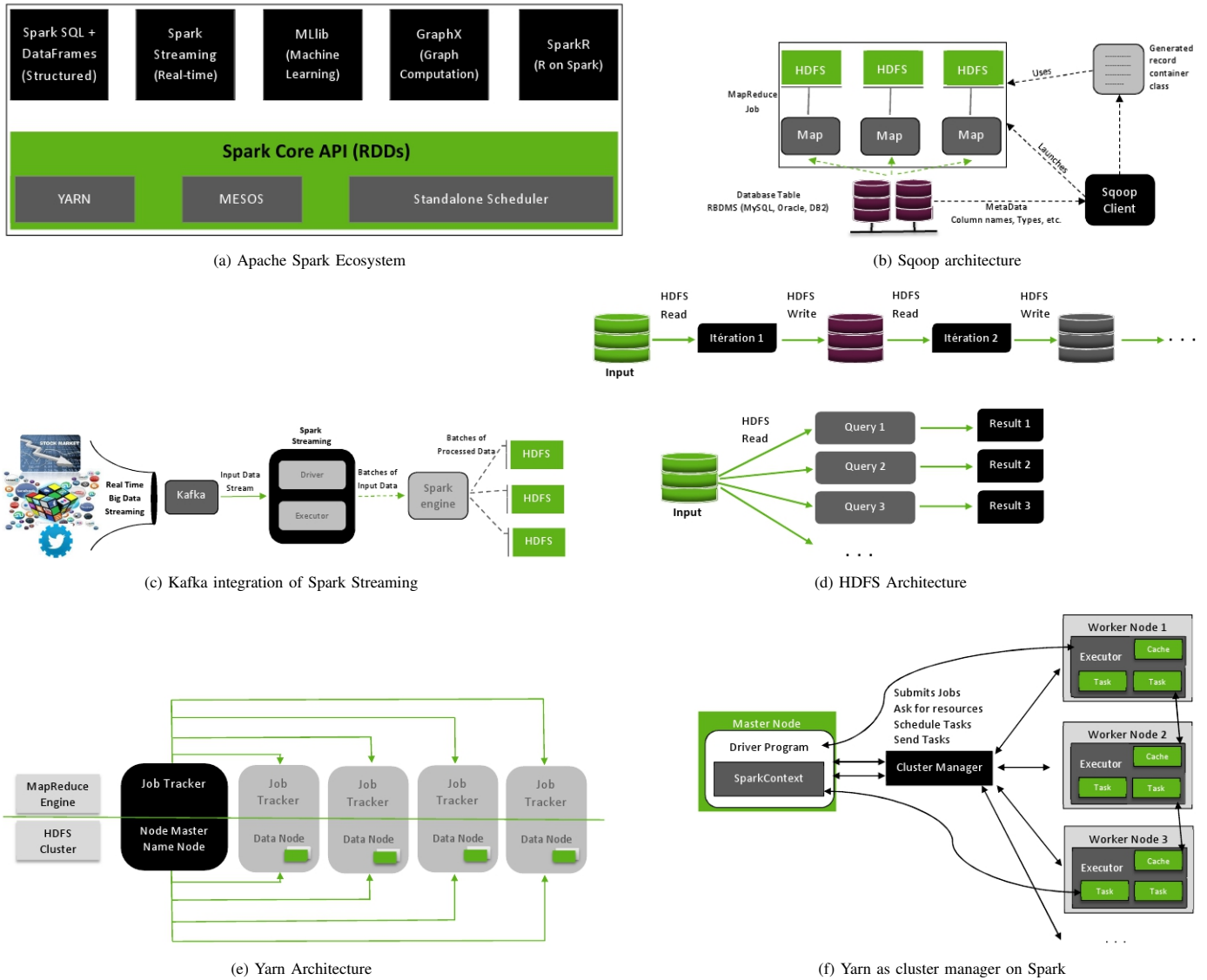


Fig. 4. Spark-based Architecture for Modeling and Big Data Analytics using ParaDist-HMM Tools

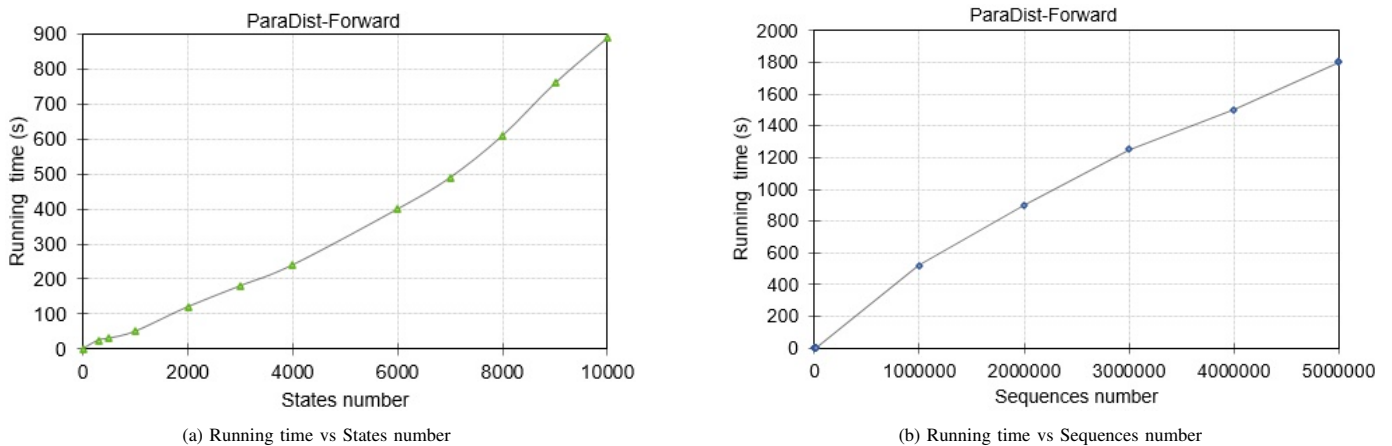


Fig. 5. ParaDist-Forward Algorithm Performances.

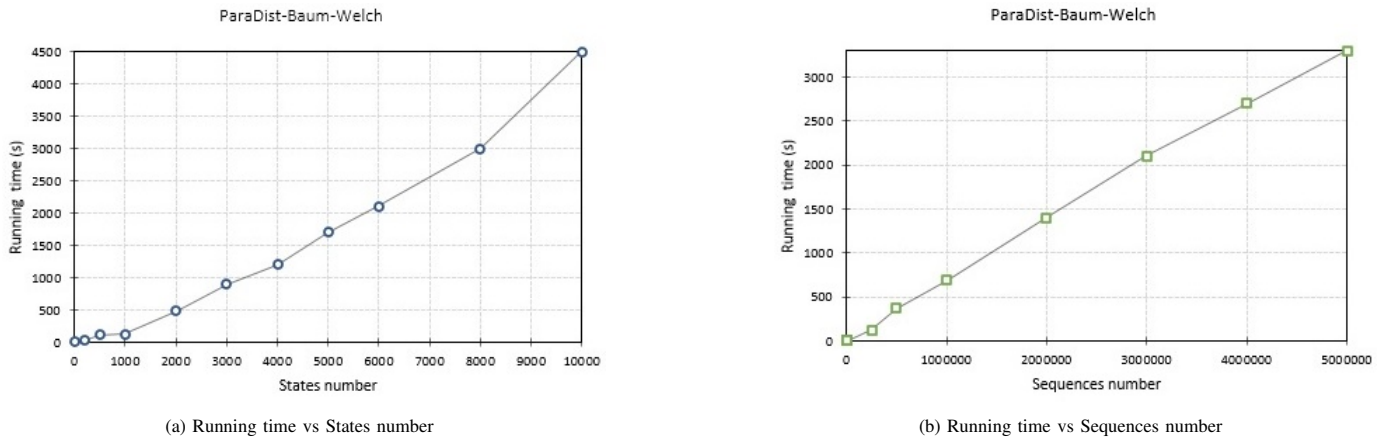


Fig. 6. ParaDist-Baum-Welch Algorithm Performances.

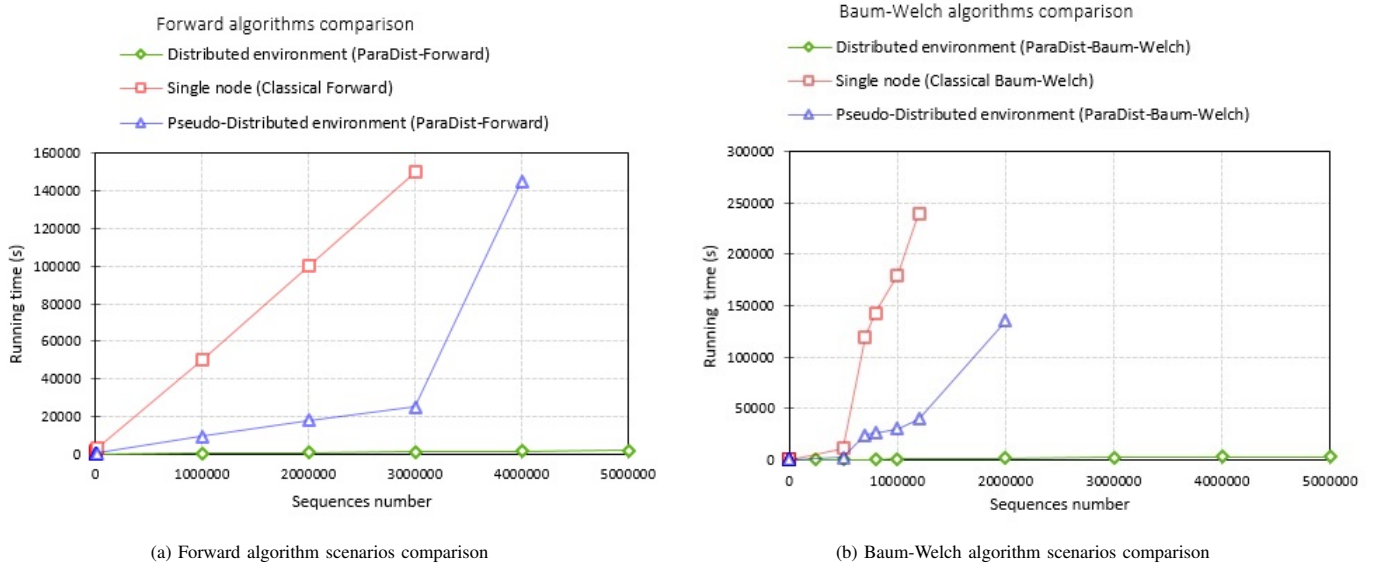


Fig. 7. Classical, Pseudo-distributed and Distributed Algorithms Comparison.

Performance Assessment of Context-aware Online Learning for Task Offloading in Vehicular Edge Computing Systems

Mutaz A. B. Al-Tarawneh¹
Computer Engineering Department
Faculty of Engineering
Mutah Univesity, Jordan

Saif E. Alnawayseh²
Electrical Engineering Department
Faculty of Engineering
Mutah Univesity, Jordan

Abstract—Vehicular Edge Computing (VEC) systems have recently become an essential computing infrastructure to support a plethora of applications entailed by smart and connected vehicles. These systems integrate the computing resources of edge and cloud servers and utilize them to execute computational tasks offloaded from various vehicular applications. However, the highly fluctuating status of VEC resources besides the varying characteristics and requirements of different application types introduce extra challenges to task offloading. Hence, this paper presents, implements and evaluates various task offloading algorithms based on the Multi-Armed Bandit (MAB) theory for VEC systems with predefined application types. These algorithms seek to make use of available contextual information to better steer task offloading. These information include application type, application characteristics, network status and server utilization. The proposed algorithms are based on having either a single MAB learner with application-dependent reward assignment, multiple application-dependent MAB learners or dedicated contextual bandits implemented as an array of incremental learning models. They have been implemented and extensively evaluated using the EdgeCloudSim simulation tool. Their performance has been assessed based on task failure rate, service time and Quality of Experience (QoE) and compared to that of recently reported algorithms. Simulation results demonstrate that the proposed contextual bandit-based algorithm outperforms its counterparts in terms of failure rate and QoE while having comparable service time values. It has achieved up to 73.4% and 21.7% average improvements in failure rate and QoE, respectively, among all application types. In addition, it efficiently utilizes the available contextual information to make appropriate offloading decisions for tasks originating from different application types achieving more balanced utilization of the available VEC resources. Ultimately, employing incremental learning to implement the proposed contextual bandit algorithm has shown a profound potential to cope with dynamic changes of the simulated VEC systems.

Keywords—Vehicular edge computing; task offloading; multi-armed bandits; contextual bandits

I. INTRODUCTION

Recently, the emergence of smart and connected vehicles has excelled the development of various types of vehicular applications such as infotainment and autonomous driving services [1], [2]. These applications are usually supported by the on-board computing and storage hardware resources. However, the ever-increasing spectrum of compute-intensive vehicular

applications and services has rendered the on-board computational resources inadequate. Hence, Vehicular Edge Computing (VEC) systems have emerged as a baseline for providing high-performance and reliable computing services for in-vehicle applications [3]. In these systems, vehicles, edge servers - instantiated at the road side units (RSUs) and cloud servers can contribute their resources to process computational tasks generated from on-board mobile devices or vehicular driving systems [4]. Hence, computational tasks within VEC systems can be offloaded to any of the available hardware resources to ensure their correct and timely execution. While task offloading can enhance task execution and improve user-perceived Quality of Service (QoS), designing an efficient task offloading scheme is not straightforward. First, the VEC environment encompasses different application classes each with different processing demands, network bandwidth requirements, timing constraints and delay sensitivity. Such diversity in application characteristics besides the unpredictable behavior of offloading requests will cause the heterogeneous computational and network resources contained in VEC infrastructure to exhibit transient and dynamic operational characteristics. These characteristics are mostly related to the utilization levels of available computational servers and the availability of network bandwidth. Second, VEC systems entail the collaboration of various entities such as vehicles, local edge servers and global cloud servers. While such a multi-component environment can lead to more versatility in task offloading, it also increases the state-space of task offloading complicating the decision to select the most appropriate entity to handle an offloaded task [5], [6]. As the dynamic changes to the VEC systems are difficult to predict or model in advance, an efficient offloading scheme should be able to learn while offloading; it should utilize its historical offloading data to steer its future offloading decisions considering both application-salient characteristics and current status of the VEC system [7]. This work targets task offloading in VEC systems with a predefined set of applications with each application having different processing, network bandwidth and timing requirements. The essence of task offloading in such systems is to enable vehicles or offloading decision makers to interact with potential offloading destinations via task offloading, learn their suitability to handle the offloaded tasks and utilize recent offloading history to guide current offloading decisions. As the set of possible offloading destinations in the considered VEC systems remain unchanged, task offloading can be formulated as a multi-armed bandit

(MAB) in which each possible offloading destination (i.e., computational server) is considered as an independent arm. Hence, pulling an arm at each round is equivalent to selecting a particular computational server to receive the offloaded task. This requires maintaining a reasonable trade-off between the exploitation (i.e., selecting the current best computational server based on past offloading decisions) and exploration (i.e., trying other servers to gain more useful and accurate information). In this regard, classical MAB solutions such as the Upper-Confidence Bound (UCB) and soft-max [8], [9], [10] become ineffective from multiple facets. On the one hand, offloading requests originate from different applications with distinct timing requirements and delay sensitivity levels hindering the process of reward formulation in the underlying MAB problem. On the other hand, the candidate arms (i.e., computational servers) may encounter dynamic changes due to their varying resource utilization levels and network connections status. To address these issues, this paper presents and evaluates three different approaches that leverage some contextual information about different application types and current status of computational servers to make offloading decisions. First, as the considered applications have different timing requirements and delay sensitivity levels, a MAB-based approach with application-dependent reward assignment is implemented and evaluated. Second, in order to ensure that an offloading decision for a particular task type is influenced only by the offloading history of similar tasks, another MAB approach, in which a dedicated bandit learner is maintained per each application type, is proposed and evaluated. Third, to cope with the dynamic and continuous changes of the VEC environment, two variations of a contextual-bandit algorithm are also proposed. This algorithm leverages incremental (online) learning to continuously adjust offloading decisions based on current environment changes. The rationale behind contextual bandits is to compute expected rewards as function of some contextual information. In order to capture variations between different arms for the same application type or variations of the same arm for different application types, this work implements contextual-bandits as an array of incremental learners with either one separate learner per arm (i.e., computational server) or one dedicated learner per each combination of arm and application type.

The rest of this paper is organized as follows. Section II discusses related research efforts. Section III presents the proposed algorithms. Section IV shows and discusses simulation results and Section V summarizes and concludes this paper.

II. RELATED WORK

Task offloading in VEC environments has recently gained a noticeable interest among researchers. Several research efforts with different decision variables and optimization goals have been proposed. In these efforts, vehicles are assumed to offload some or all of their tasks using vehicles to everything (V2X) communication technologies. Typically, V2X is a general term that indicates different communication models used by the vehicles to offload their tasks. In this context, Vehicle to Vehicle (V2V), Vehicle to RSU (V2R), Vehicle to Pedestrian (V2P) and Vehicle to Infrastructure (V2I) can be utilized [11]. Sun et. al. [12] have proposed an adaptive learning based task offloading (ALTO) algorithm for the dynamic VEC systems. They have proposed a MAB-based solution that works in a

distributed manner and targets minimizing the average delay of task offloading. However, their proposed algorithm focused on V2V task offloading. On the other hand, Zhang et. al. [7] have formulated task offloading as a mortal MAB problem in which tasks can be offloaded to neighboring edge nodes. While contextual information obtained from various edge nodes were considered when making an offloading decision, the presence of different applications with distinct characteristics and timing requirements was not considered. On the other hand, Xu et. al. [13] have formulated task offloading in VEC environments as a multi-objective optimization problem. They have solved the optimization problem using genetic algorithm with the goal of minimizing offloading latency and improving resource utilization. In their proposed method, tasks can be offloaded either to edge servers or other vehicles. However, since task offloading is an online problem and its constituent task characteristics and environment dynamics are not known in advance, finding an offline task offloading solution may not be effective in real-life.

Dai et al. [14] have formulated offloading destination selection and load balancing in VEC systems as a mixed-integer nonlinear programming problem. They have proposed an approximation heuristic algorithm to solve this problem. Their proposed algorithm is assumed to run on the vehicles in a distributed manner. In addition, they have assumed that some parts of the tasks can be executed locally using in-vehicle resources while the rest can be offloaded to the VEC server. However, dividing task execution into several parts and then combining the results is error-prone and may not be suitable for delay-sensitive applications such as accident prevention services.

Wang et al. [15] have employed a game theory-based technique to find the offloading probability of each vehicle in the VEC system. Their primary goal was to maximize the utility of each vehicle. The vehicles adjust their offloading probabilities by considering the offloading probability of other vehicles in the previous stage. Based on the computed probabilities, tasks can be executed locally, or offloaded to the edge server. However, they did not consider task offloading to the global cloud server neither did they consider the presence of applications with different requirements.

Liu et al. [16] have utilized a matching-based approach for minimizing the network delay associated with task offloading. In their work, the VEC system is composed of three layers that include the vehicles, RSUs and a macro base station (MBS). The MBS is responsible for performing task offloading and handover operations. Hence, all the vehicles and RSUs are assumed to be connected to the MBS. The matching algorithm operates iteratively based on matching requests sent from the vehicles to the RSUs. However, their work was based on a fixed-latency network model in which the latency of the wide-area network (WAN) is assumed to be fixed. Hence, their work did not consider the impact of network status on task offloading especially that the matching requests will create extra load on the available network bandwidth.

Feng et al. [17] have proposed a hybrid vehicular cloud (HVC) framework to increase the computing capacity of vehicles by utilizing computational resources of other neighboring vehicles, RSUs and the cloud. The goal of their proposed online algorithm is to increase the number of successfully offloaded and executed tasks while minimizing cellular net-

work usage. Their proposed algorithm seeks to first find the idle slots on other neighboring vehicles and RSUs considering both the estimated transmission and execution delays. If no idle slots are found on the neighbouring vehicles or the RSUs, the cellular network is used to access the cloud. All devices in the VEC system are assumed to work collaboratively by broadcasting a beacon message. Computational tasks are scheduled consecutively based on their anticipated transmission and processing delays. However, incorrect or misleading information provided by some malicious vehicles may cause some tasks to fail. On the other hand, Jiang et. al. [18] have introduced task replication technique to improve service reliability in VEC systems. In their proposed approach, task replicas can be simultaneously offloaded to multiple vehicles to be processed. However, one drawback of their proposed framework is that it needs frequent state information update and can place significant overhead on the network bandwidth.

Sonmez et al. [6] have recently proposed a machine learning-based task offloading scheme for VEC systems. They have considered a multi-access, multi-tier VEC architecture that consists of three main layers, namely, the vehicles, RSUs and cloud servers. In addition, they have also assumed a multi-access communication framework in which vehicular wireless local area network (WLAN), wide-area network (WAN) and cellular network can be used for V2I task offloading. Their task offloading scheme is based on a two-stage process in which dedicated regression and classification models are maintained per each potential offloading destination. During the first stage, the classification models are consulted to predict which devices could successfully handle the offloaded task. In the second stage, the regression models are employed to predict the time required to execute the offloaded task (i.e., service time) on each of the devices identified during the first stage. Thereafter, the device with the lowest predicted service time is chosen to receive the offloaded task. However, their work is based on having a static dataset to train the regression and classification models. However, such a static dataset may not be available in real-life as the VEC environment from which the data is collected changes continuously. In addition, their used regression and classification models remain static and do not acquire any new knowledge from the dynamic changes of VEC environment. In other words, when the VEC environment conditions to which the static models are exposed differ from those used for model training, their proposed offloading scheme may fall short and lead to poor performance.

In this work, three different online MAB-based task offloading schemes are implemented and evaluated based on the VEC architecture presented in [6]. The common theme among these schemes is to account for the presence of applications with different requirements and dynamically adjust the offloading decisions based on the dynamic conditions of the VEC system.

III. ONLINE VEHICULAR TASK OFFLOADING

As task offloading requests are sequentially generated in a dynamic manner, task offloading becomes an online sequential decision making process that cannot be handled using traditional offline optimization tools. Instead, it can be formulated and solved using MAB theory. Hence, this work implements several MAB-based task offloading algorithms. In addition,

it evaluates their performance in terms of the percentage of satisfied task offloading requests, task response time and QoE - under dynamically changing server utilization levels and network conditions.

A. VEC System Overview

A typical VEC system is composed of multiple edge servers augmented by the global cloud resources. In addition, the underlying communication infrastructure encompasses multiple technologies such as WLAN, MAN and WAN [19]. Such a heterogeneous architecture with time-varying offloading patterns leads to a dynamic scene that requires proper management of task execution. In this regard, the task offloading engine tries to preserve an efficient operation of the entire VEC system by selecting the best available computational server to receive an offloaded task. The decision on which server to choose is substantially demanding as it should consider both task characteristics, computational server utilization and network status. This work assumes a multi-tier and multi-access VEC system's architecture in which both local edge servers and global cloud servers can receive offloaded tasks [6]. In this architecture, vehicles can offload their computational tasks to the edge servers (i.e., V2R) or to the cloud servers (i.e., V2I). On the one hand, tasks can be offloaded to the edge servers using a short-range WLAN communication protocol such as the IEEE 802.11 used in [20], [21]. On the other hand, vehicles can offload their tasks to the cloud servers using the Internet connection (WAN), which provides a more flexible and high-bandwidth network interface. Similar to the model proposed in [6], vehicular tasks can be offloaded to the cloud either through the serving RSUs, which are assumed to use fiber connection to the cloud, or using the cellular network's broadband connection. Furthermore, the RSUs in the considered VEC architecture are also connected through a Metropolitan Area Network (MAN). This allows RSUs to form a shared resource pool in which task migration can be performed to handle the handover problem as proposed in [16]. In the assumed handover scheme, when a vehicle leaves the range of its current serving RSU before the results of the offloaded task are received, those results are transmitted to that vehicle in a multi-hop manner via the other RSUs in the VEC system. The handover process only fails if the offloading vehicle leaves the range of its current serving RSU while uploading a task or downloading a result.

Therefore, the considered VEC system allows vehicles to offload their tasks either to the edge sever, cloud server through RSU or cloud server through cellular network.

B. Task Offloading Algorithms

In this work, the considered algorithms are based on the MAB theory which is a Reinforcement Learning (RL) approach to maximize the total cumulative reward through sequential decision making. As shown in Fig. 1, a typical RL problem is modelled as an environment whose state is continuously observed by an agent. As shown, the agent observes the environment state (S1) and takes an action (A). Consequently, the environment responds by transitioning to state (S2) and sending a reward (R) to the agent. The reward may be positive or negative. Over a series of such trials and

errors, the agent learns an optimal policy (i.e., a mapping from states to actions) to maximize the long-term reward.

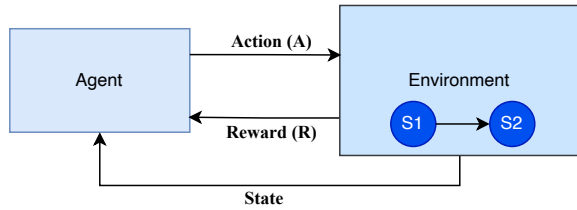


Fig. 1. Reinforcement Learning Flow.

In reality, the RL problem can be simply abstracted as a MAB problem. As shown in Fig. 2. MAB problems do not account for the environments and their state changes. In other words, an agent observes only the actions it takes and the associated rewards it receives and tries to compose the optimal strategy accordingly. The rationale behind solving MAB problems is to try and explore the actions involved in the action space and realize the unknown distributions of the rewards. Therefore, in MAB problems, the agents will ultimately try different actions and maintain a trade-off between exploration and exploitation to devise the optimal policy.

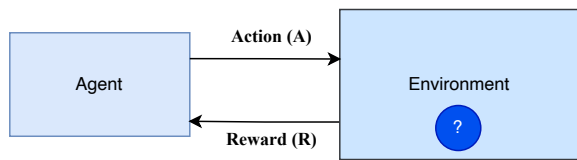


Fig. 2. Multi-armed Bandit Problem Flow.

Evidently, the main drawback of MABs is that the agents totally ignore the environment state when making an action. The environment state can provide significantly useful insights that can help the agent in devising an efficient policy much faster. Utilizing some useful elements of the environment state has introduced a new class of algorithms know as contextual or context-aware bandits [22], [23], [24], illustrated in Fig. 3. Here, instead of managing the trade-off between exploration and exploitation randomly, the agent obtains some context (i.e., contextual information) about the environment and utilizes that information to properly manage the actions. The notion of context is different from that of the state used in the RL problem formulation. A context is simply some useful knowledge about the environment that helps the agent take a proper action. For example, in the case of task offloading, the context may provide some information about the application type to which an offloaded task belongs.

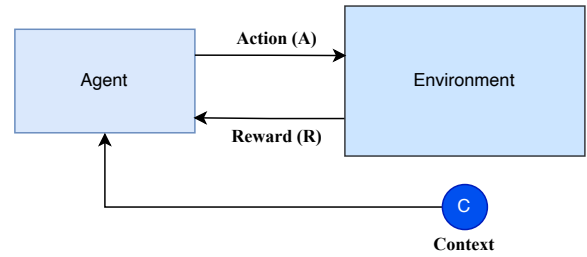


Fig. 3. Context-aware Multi-armed Bandit Problem Flow.

In the task offloading problem - formulated as a MAB or contextual MAB problem, the task offloading engine is the agent, the environment represents the VEC system while the reward is a numeric representation of user's perception of the quality of the offloading service. On the other hand, the action space consists of all offloading options i.e., offloading to the edge servers, offloading to the cloud servers via the RSU or offloading to the cloud via the cellular network. The following sections (i.e., III-B1, III-B2 and III-B3) describe each of the implemented context-aware task offloading algorithms.

1) Task offloading with application-dependent rewards:

This section explains the implemented MAB-based task offloading algorithms in which the reward assigned to the agent is application-dependent. The reward assigned after receiving the results of an offloaded task is computed based on the observed task's response time, maximum tolerable delay of the application to which the offloaded task belongs and that application's delay sensitivity. The response time of an offloaded task can be computed as shown in equation 1.

$$\begin{aligned}
 T_R &= t_u + t_p + t_d \\
 t_u &= \frac{T_s + T_{IN}}{Network_{ub}} \\
 t_p &= \frac{T_{IC}}{Server_{MIPS}} \\
 t_d &= \frac{T_{OUT}}{Network_{db}}
 \end{aligned} \tag{1}$$

Where T_R is the total response time of the offloaded task in seconds, t_u is the time required to upload the task and its input file to the selected server, t_p is the execution time of the offloaded task, t_d is the time required to download the results to the offloading vehicle, T_s is the size of the offloaded task's binary in Megabyte (MB), T_{IN} is the task's input file size in MB, $Network_{ub}$ the uplink bandwidth of the network connection associated with the selected arm (i.e., computational server) in MB/s, T_{IC} represents the instruction count of the offloaded task, $Server_{MIPS}$ is the processing capacity of the associated server in million instructions per second (MIPS), T_{OUT} is the task's output file size and $Network_{db}$ is the downlink bandwidth of the used network connection.

Assuming that the response time observed after the agent has offloaded a task (i), generated from an application (A) whose maximum delay requirement is T_{max} and delay sensitivity is α_A , is T_i . Then, the reward assigned to the agent (R_i) is computed as shown in equation 2. This formulation is based on the notion of Quality of Experience (QoE) proposed in [6].

$$R_i = \begin{cases} 0, & \text{if } i \text{ has failed} \\ 0, & \text{if } T_i \geq 2T_{max} \\ (1 - \frac{T_i - T_{max}}{T_{max}}) \cdot (1 - \alpha_A), & \text{if } T_{max} \leq T_i < 2T_{max} \\ \alpha_A \cdot R_{max}, & \text{if } T_i \leq T_{max} \end{cases} \quad (2)$$

Where R_{max} is the maximum possible reward, $\alpha_A \in [0, 1]$ and $1 - \alpha_A$ refers to the delay tolerance of the associated application. In other words, a high value of α_A indicates that the associated application is a delay-sensitive application while a low value of α_A represents a delay-tolerable application. Apparently, the value of the reward is directly linked to application characteristics (i.e., the maximum delay requirement and delay sensitivity). Hence, two similar response time values obtained for two different applications - with distinct requirements will be viewed differently by the agent. Consequently, the agent will reasonably scale the cumulative reward associated with a particular arm in proportion to application characteristics; while a selected arm (i.e., computational server) might be suitable for a particular application type, it may not satisfy the requirements of other application types.

In a MAB problem with k possible arms, there are k possible actions i.e., arm selection choices. Each action has an expected reward provided that the action is selected. This expected reward is known as the value of that action and denoted as $q_*(a)$. The action selected at time instant t is denoted as A_t . Hence, the value of an arbitrary action a , is the expected reward given that a is selected by the agent, as shown in equation 3 [25].

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a] \quad (3)$$

If the agent knew the value associated with each action, then it would be trivial to solve the MAB problem: the agent would always select the action with highest value. However, the agent does not know the action values with certainty, although it may have estimates. The estimated value of action a at time instance t is denoted as $Q_t(a)$. If the agent maintains estimates of the values of different actions, then at any time instant there exist at least one action whose estimated value is the highest. This action - with the highest estimated value is known as the greedy action. Hence, a simple action selection policy is to always pick the greedy action, as given in equation 4 [25].

$$A_t \doteq \underset{a}{\operatorname{argmax}} Q_t(a) \quad (4)$$

Where the *argmax* operator returns the action that maximizes the enclosed expression. If the agent selects a greedy action, the agent is said to be exploiting its current knowledge of the values of the actions. If instead the agent picks one of the non-greedy actions, then the agent is said to be exploring, because this enables the agent improve its estimates of the values of the non-greedy actions. While exploitation is the right thing that the agent can do to maximize the expected reward on the one step, exploration may yield greater total reward in the long run. For example, suppose the value of the greedy action is known with certainty, while some other actions are

anticipated to be nearly as good but with some high degree of uncertainty. The uncertainty is such that at least one of the other actions is probably better than the greedy action, but the agent does not know which one. If the agent has many time steps ahead on which to choose among actions, then it may be better to explore the non-greedy actions and identify which of them are better than the greedy action. Because the agent is not able to both explore and exploit with any single action selection, the conflict or trade-off between exploration and exploitation should be properly addressed. This work considers two possible MAB algorithms that handle the exploitation-exploration dilemma taking into account the uncertainty in the estimates of action values. These algorithms are the Upper-Confidence Bound (UCB) and the soft-max bandit algorithms [8], [9], [10]. The UCB action selection policy works based on the premise that it would be better to choose from the non-greedy actions in accordance with their potential for actually being optimal, considering both how close their estimates are to being maximal besides the uncertainties in those estimates. This action selection policy is given in equation 5 [8], [26].

$$A_t \doteq \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c \sqrt{\frac{\ln(N)}{N_t(a)}} \right] \quad (5)$$

Where N is the number of action selections performed by the agent, $N_t(a)$ is the number of times action a has been selected so far and $c > 0$ is the exploration control parameter. The rationale behind the UCB action selection is that the square root term is a measure of the uncertainty in the estimate of the value of action a . Hence, the quantity being max'ed over is therefore a kind of upper bound on the potential true value of action a , with parameter c determining the confidence level. When action a is selected by the agent, $N_t(a)$ increases and its associated uncertainty is reduced, and, since $N_t(a)$ appears in the denominator, the uncertainty term decreases as well. On the other hand, every time an action other than a is selected by the agent, the value of t increases but $N_t(a)$ does not. Hence, as t appears in the numerator, the estimate of uncertainty - associated with a increases. In addition, the use of the natural logarithm indicates that the increases in uncertainty get smaller over time, but are unbounded; all actions will ultimately be selected. However, actions with lower value estimates, or that have already been selected more often, will be selected by the agent with decreasing frequency over time.

On the other hand, the soft-max algorithm picks each action with a probability that is proportional to its current estimated value $Q_t(a)$ as shown in equation 6 [25], [26].

$$Pr\{A_t = a\} \doteq \frac{e^{Q_t(a)/\tau}}{\sum_{j=1}^k e^{Q_t(j)/\tau}} \quad (6)$$

Where τ is a temperature parameter used to control the randomness of action selection. When $\tau = 0$, the algorithm acts greedily. When τ increases to infinity, the algorithm will select actions uniformly at random. In other words, the soft-max algorithm learns a numerical preference for each action a , which is proportional to the action value (i.e., $Q_t(a)$). The larger the preference, the more frequently that action is selected.

In the two algorithms, after an action a is selected and a reward is received, the estimated value of that action is incrementally updated as shown in equation 7 [25].

$$Q_{t+1}(a) = Q_t(a) + \frac{1}{N_a(t)} [R_t - Q_t(a)] \quad (7)$$

Where $Q_{t+1}(a)$ is the new estimate of the action value, $Q_t(a)$ is the old estimate of that action's value, R_t (equation 2) is the recently received reward.

Algorithms 1 and 2 show high-level pseudo-codes of the UCB and soft-max algorithms, respectively. As shown, the two algorithms keep track of dynamically changing variables such as the number of times each arm/action has been selected (i.e., the N_a array) and the action value estimates (i.e., the Q array). On the other hand, each algorithm implements three basic functions, namely, *initialize*, *chooseArm* and *updateArmValue*. The *initialize* procedure is used to initialize algorithm's variables and data structures. The *chooseArm* procedure is responsible for arm selection while the *updateArmValue* is used to update the selected action's value after receiving the reward from the environment. While the *initialize* and *chooseArm* functions are similar in the two algorithms, the *chooseArm* procedures are different. In the UCB algorithm, the *chooseARM* procedure computes the UCB values of different actions based on their current value estimates and their uncertainty measures. Thereafter, it acts greedily on the computed UCB values. On the other hand, the *chooseArm* procedure in the soft-max algorithm computes action probabilities - in proportion to their current value estimates and performs categorical draw to select actions based on their computed probabilities. Algorithm 3 shows a high-level abstraction of the main functions involved in the proposed MAB-based task offloading algorithm with application dependent reward assignment. First, the task offloading agent is initialized as either a UCB or soft-max algorithm (line 1). The utilized MAB algorithm is configured to have three possible actions, namely, offloading to the edge server (Edge), offloading to the cloud server via the RSU (cloudRSU) and offloading to the cloud server via the cellular network (cloudCN), as shown in lines 2 and 3 of algorithm 3. Every time an offloading request is received by the offloading agent, the *selectOffloadingDestination* procedure (lines 4-9) is invoked. This procedure will call the associated MAB algorithm to select a particular server to handle the offloaded task (lines 5-6), update the task's meta-data to maintain an association between the task and the selected server (line 7) and then offload the task to the selected server (line 8). When the results of the offloaded task are returned from selected server, the *taskCompleted* procedure (lines 10-22) will be called. This procedure will first check the completed task's meta-data to determine the server that has handled the task (lines 11-20) and then ask the MAB learner (i.e., algorithm) to update the value of that server based on the obtained reward (line 21). On the other hand, if the offloaded task fails, the *taskFailed* procedure (lines 23-35) can be called to update the value of the associated server accordingly. As shown, the *taskCompleted* and *taskFailed* procedures will eventually call the *updateArmValue* procedure defined in algorithms 1 and 2.

Algorithm 1 UCB Algorithm

Input: *Task*

Output: *selectedArm*

```
1:  $numArms \leftarrow$  number of arms
2:  $N_a[numArms] \leftarrow$  array of individual arm pulls
3:  $Q[numArms] \leftarrow$  array of action/arm values
4: procedure INITIALIZE( $n$ )
5:    $numArms \leftarrow n$ 
6:    $N \leftarrow 0$ 
7:   for  $i \leftarrow 1$  to  $numArms - 1$  do
8:      $N_a[i] \leftarrow 0$ 
9:      $Q[i] \leftarrow 0$ 
10:  end for
11: end procedure
12: procedure CHOOSEARM()
13:    $N \leftarrow 0$ 
14:   for  $i \leftarrow 0$  to  $numArms - 1$  do
15:      $count \leftarrow N_a[i]$ 
16:     if  $count = 0$  then
17:        $N_a[i] \leftarrow 1$ 
18:       return  $i$ 
19:     end if
20:      $N \leftarrow N + N_a[i]$ 
21:   end for
22:    $ucbQ[numArms] \leftarrow$  temporary array of UCB values
23:   for  $i \leftarrow 0$  to  $numArms - 1$  do
24:      $ucbQ[i] \leftarrow Q[i] + \sqrt{\frac{2 * \ln(N)}{N_a[i]}}$ 
25:   end for
26:    $selectedArm \leftarrow 0$ 
27:   for  $i \leftarrow 1$  to  $numArms - 1$  do
28:      $newValue \leftarrow ucbQ[i]$ 
29:     if  $newValue > ucbQ[selectedArm]$  then
30:        $selectedArm \leftarrow i$ 
31:     end if
32:   end for
33:    $N_a[selectedArm] \leftarrow N_a[selectedArm] + 1$ 
34:   return  $SelectedArm$ 
35: end procedure
36: procedure UPDATEARMVALUE( $arm, task, success$ )
37:   if  $success = False$  then
38:      $reward \leftarrow 0$ 
39:   else
40:      $\alpha_A \leftarrow task.delaySensitivity$ 
41:      $T_i \leftarrow task.responseTime$ 
42:      $T_{max} \leftarrow task.maxDelayRequirement$ 
43:      $R_{max} \leftarrow 1$ 
44:      $reward \leftarrow$  result of equation 2
45:   end if
46:    $Q[arm] = Q(arm) + \frac{1}{N_a(arm)} [reward - Q(arm)]$ 
47: end procedure
```

2) *Task offloading with application-dependent bandits:*

This section presents the application-dependent MAB-based task offloading algorithm. The basic idea behind this algorithm is to ensure that the offloading decision for a particular task is influenced by the offloading history of similar tasks i.e., tasks originating from similar application type. Algorithm 4 gives a pseudo-code of the application-dependent task offloading algorithm. As shown, the algorithm proceeds (lines 3-6) by

Algorithm 2 Soft-max Algorithm

Input: *task*

Output: *selectedArm*

```
1: numArms  $\leftarrow$  number of arms
2:  $N_a[\textit{numArms}] \leftarrow$  array of individual arm pulls
3:  $Q[\textit{numArms}] \leftarrow$  array of action/arm values
4:  $\tau \leftarrow$  temperature value
5: procedure INITIALIZE(n)
6:   numArms  $\leftarrow n$ 
7:    $N \leftarrow 0$ 
8:   for  $i \leftarrow 1$  to numArms - 1 do
9:      $N_a[i] \leftarrow 0$ 
10:     $Q[i] \leftarrow 0$ 
11:   end for
12: end procedure
13: procedure CHOOSEARM()
14:   sumQ  $\leftarrow 0$ 
15:   for  $i \leftarrow 0$  to numArms - 1 do
16:     sumQ  $\leftarrow$  sumQ +  $e^{Q(i)/\tau}$ 
17:   end for
18:   probabilities[numArms]  $\leftarrow$  temporary array of soft-
max probabilities
19:   for  $i \leftarrow 0$  to numArms - 1 do
20:     probabilities[ $i$ ]  $\leftarrow \frac{e^{Q(i)/\tau}}{\textit{sumQ}}$ 
21:   end for
22:   return categoricalDraw(probabilities)
23: end procedure
24: procedure CATEGORICALDRAW(probabilities)
25:   rand  $\leftarrow$  random double  $\in [0, 1]$ 
26:   cumulativeP  $\leftarrow 0$ .  $\triangleright$  cumulative probability
27:   for  $i \leftarrow 0$  to numArms - 1 do
28:     cumulativeP  $\leftarrow$  cumulativeP + probabilities[ $i$ ]
29:     if cumulativeP > rand then
30:       return  $i$ 
31:     end if
32:   end for
33:   return numArms - 1
34: end procedure
35: procedure UPDATEARMVALUE(arm, task, success)
36:   if success = False then
37:     reward  $\leftarrow 0$ 
38:   else
39:      $\alpha_A \leftarrow$  task.delaySensitivity
40:      $T_i \leftarrow$  task.responseTime
41:      $T_{max} \leftarrow$  task.maxDelayRequirement
42:      $R_{max} \leftarrow 1$ 
43:     reward  $\leftarrow$  result of equation 2
44:   end if
45:    $Q[\textit{arm}] = Q(\textit{arm}) + \frac{1}{N_a(\textit{arm})} [\textit{reward} - Q(\textit{arm})]$ 
46: end procedure
```

initializing each offloading engine, associated with each application type, as a 3-arm MAB. This MAB can be either a UCB or a soft-max algorithm. In other words, the offloading agent maintains a separate offloading engine for each application type.

As shown, algorithm 4 implements three main procedures. The *selectOffloadingDestination* procedure (lines 7-13) - used for server selection will first identify the application type

Algorithm 3 Task Offloading Algorithm with Application-dependent Rewards

Input: *task*

Output: *Selected Offloading Destination*

```
1: taskOfFloder  $\leftarrow$  MAB
2: servers [] = {Edge, cloudRSU, cloudCN}
3: taskOfFloder.Initialize( $n = 3$ )
4: procedure SELECTOFFLOADINGDESTINATION(task)
5:   arm  $\leftarrow$  taskOfFloder.chooseARM()
6:   server  $\leftarrow$  servers[arm]
7:   task.setAssociatedServer(server)
8:   offload task to server
9: end procedure
10: procedure TASKCOMPLETED(task)
11:   server  $\leftarrow$  task.getAssociatedServer()
12:   if server = Edge then
13:     arm = 0
14:   end if
15:   if server = cloudRSU then
16:     arm = 1
17:   end if
18:   if server = cloudCN then
19:     arm = 2
20:   end if
21:   taskOfFloder.updateArmValue(arm, task, True)
22: end procedure
23: procedure TASKFAILED(task)
24:   server  $\leftarrow$  task.getAssociatedServer()
25:   if server = Edge then
26:     arm = 0
27:   end if
28:   if server = cloudRSU then
29:     arm = 1
30:   end if
31:   if server = cloudCN then
32:     arm = 2
33:   end if
34:   taskOfFloder.updateArmValue(arm, task, False)
35: end procedure
```

(line 8), utilize the associated MAB learner to select a particular server (lines 9-10), record the task-to-server association (line 11) and then offload the task to the selected server (line 12). On the other hand, the *taskCompleted* procedure (lines 14-27) - invoked upon the receipt of the offloaded task's results will use that task's type (line 15) and other meta-data (lines 16-25) to ask the related MAB learner to update the value of the server to which the task was offloaded (line 26). In addition, the *taskFailed* procedure (lines 28-41) is responsible for handling task failure; it identifies the task type and that task's associated server (lines 29-39) and updates the server value accordingly (line 40). As algorithm 4 defines a dedicated MAB learner for each application type, the *updateArmValue* procedure called in lines 26 and 40 is redefined to compute the value of the reward as shown in equation 8 instead of equation 2; tasks with the same application type are assumed to have the same value of delay sensitivity (α_A).

Algorithm 4 Task Offloading Algorithm with Application-dependent MAB Learners

Input: *task*

Output: Selected Offloading Destination

```

1: numApps ← Number of application types
2: servers [] = {Edge, cloudRSU, cloudCN}
3: for i ← 0 to numApps - 1 do
4:   taskOfFloder[i] ← MAB
5:   taskOfFloder[i].Initialize(n = 3)
6: end for
7: procedure SELECTOFFLOADINGDESTINATION(task)
8:   type ← task.getApplicationType()
9:   arm ← taskOfFloder[type].chooseARM()
10:  server ← servers[arm]
11:  task.setAssociatedServer(server)
12:  offload task to server
13: end procedure
14: procedure TASKCOMPLETED(task)
15:  t ← task.getApplicationType()
16:  server ← task.getAssociatedServer()
17:  if server = Edge then
18:    arm = 0
19:  end if
20:  if server = cloudRSU then
21:    arm = 1
22:  end if
23:  if server = cloudCN then
24:    arm = 2
25:  end if
26:  taskOfFloder[t].updateArmValue(arm, task, True)
27: end procedure
28: procedure TASKFAILED(task)
29:  t ← task.getApplicationType()
30:  server ← task.getAssociatedServer()
31:  if server = Edge then
32:    arm = 0
33:  end if
34:  if server = cloudRSU then
35:    arm = 1
36:  end if
37:  if server = cloudCN then
38:    arm = 2
39:  end if
40:  taskOfFloder[t].updateArmValue(arm, task, False)
41: end procedure

```

$$R_i = \begin{cases} 0, & \text{if } i \text{ has failed} \\ 0, & \text{if } T_i \geq 2T_{max} \\ (1 - \frac{T_i - T_{max}}{T_{max}}), & \text{if } T_{max} \leq T_i < 2T_{max} \\ 1, & \text{if } T_i \leq T_{max} \end{cases} \quad (8)$$

3) *Task offloading with incremental learning:* This section presents two variations of an algorithm in which incremental (i.e., online) learning is used to guide task offloading agents. The presented algorithm is inspired by the idea of contextual bandits used in some domains such as recommendation systems [27], [28]. As shown in [22], the main principle in contextual bandits is to construct a linear model that

can be used to predict the expected reward of choosing a particular action considering some contextual information. The parameters of this model are continuously updated based on the true observed reward. Hence, this work presents a new algorithm that employs the idea of contextual bandits for task offloading. The rationale behind this algorithm is to develop and maintain an online model that predicts a task response time based on contextual information such as task processing requirements, server utilization and network latency, as shown in equation 9 [22].

$$\mathbb{E}[T_t | \mathbf{x}_t] = f_{\theta}(\mathbf{x}_t) = \mathbf{x}_t^T \theta = \sum_{j=0}^n \theta_j x_{tj} \quad (9)$$

Where T_t is predicted response time, \mathbf{x}_t is the context vector at time t and θ is the model coefficients vector and n is the number of parameters in the context vector. The vector \mathbf{x}_t contains contextual information related to application characterises and the VEC environment status. Hence, the goal of online learning process is to find the coefficients vector θ that would minimize the error between the predicted response time (i.e. T_t) - computed before task offloading and the actual response time (T_{a_t}) observed after offloading. In other words, the learning process seeks to find the values of θ that would minimize a particular cost function C_{θ} . As the model given in equation 9 represents a liner regression model, the squared loss function given in equation 10 is a suitable choice for the cost function [29]. Evidently, this function computes the squared error or difference between the predicted and observed response times. The average cost per training instance can be computed as given in equation 11, which computes the mean squared error (MSE) [29].

$$C_{\theta} = \frac{1}{2} (f_{\theta}(\mathbf{x}_t^{(i)}) - T_{a_t}^{(i)})^2 \quad (10)$$

$$MSE_{\theta} = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (f_{\theta}(\mathbf{x}_t^{(i)}) - T_{a_t}^{(i)})^2 \quad (11)$$

Where m is the number of training instances. Hence, the goal of model training is to find the values of θ that would minimize the MSE over the whole training set. In traditional machine learning settings, the values of θ are usually computed offline using some optimization techniques such the gradient decent - assuming the presence of a static training dataset. However, in the task offloading problem, a static training dataset of offloading requests does not usually exist; offloading requests appear as a dynamic stream of instances. Hence, the model parameters (i.e., values of θ) should be incrementally learned and adjusted based on the incoming offloading requests. In this regard, stochastic gradient decent (SGD) provides a viable tool to incrementally adjust the values of θ based on individual training instances [30]. It can perform a parameter update for each training instance ($\mathbf{x}_t^{(i)}$, $T_{a_t}^{(i)}$). After each training instance, the values of θ will be updated as shown in equation 12.

$$\theta_j = \theta_j - \eta \nabla_{\theta_j} C_{\theta}(\mathbf{x}_t^{(i)}, T_{a_t}^{(i)}) \quad (12)$$

Where η is the learning rate and ∇_{θ_j} is the gradient of the cost function with respect to θ_j . Hence, this work employs SGD to dynamically fit and adjust linear models for response time prediction. Apparently, as the VEC environment contains different offloading options i.e., servers, a single linear model would not suffice for all possible servers. Hence, the first variant of the SGD-based offloading algorithm maintains three separate linear models for response time predictions; a model for edge server (SGD_{edge}) and two other models for cloud via RSU (SGD_{cRSU}) and cloud via CN (SGD_{cCN}), respectively. Table I summarizes the contextual features used by each model. Each model uses its own context vector to make response time predictions. While different context vectors contain some similar application characteristics such as the task's instruction count, each vector contains server-specific features such as that server's utilization level and the upload and download latencies of the associated network connection. It is worth noting that the SGD_{edge} , unlike cloud-related models, uses the current utilization of the edge server to predict response time. In general, edge servers are not as resource-enriched as cloud servers and their utilization levels could have a significant impact on the response time.

TABLE I. CONTEXTUAL FEATURES OF SGD-BASED MODELS

Model	Contextual features
SGD_{edge}	Task instruction count (T_{IC}), edge server utilization (U_e), WLAN upload latency ($WLAN_u$), WLAN download latency ($WLAN_d$)
SGD_{cRSU}	Task instruction count (T_{IC}), WAN upload latency (WAN_u), WAN download latency (WAN_d)
SGD_{cCN}	Task instruction count (T_{IC}), CN upload latency (CN_u), CN download latency (CN_d)

Algorithms 5 and 6 show pseudo-codes of the main parts of the incremental learning-based task offloading algorithm. These algorithms follow the notation of the WEKA API used for implementing the algorithm in the used simulation tool [31]. As show in algorithm 5, the algorithm initializes all models as an SGD-based linear models (lines 1-3). These models are initialized with arbitrary values of the model coefficients θ .

In order to allow the constructed models to make educated predictions of response time, the algorithm will first utilize a round-robin-based offloading until a batch of instances, for each server, with a predefined size is obtained. Each instance of the batch records the server's contextual features (Table I) - at the time of offloading besides the observed response time value under that context. Thereafter, a model, for each server, is trained on the associated batch (lines 4-6). When an offloading request is received, the *selectOffloadingDestination* procedure (lines 7-15) is invoked. This procedure will first observe the context associated with each possible offloading option and call the constructed models to make response time predictions, with a prediction per each offloading option (lines 8-10). Then, the offloading option with the least predicted response time is chosen for task offloading. In addition, the incremental learning-based algorithm maintains a dictionary to keep track of the selected server's context at the time of offloading (lines 12-13). This dictionary will later be used for updating the respective model parameters when the result of offloading is disclosed.

On the other hand, when the results of offloading are suc-

Algorithm 5 Task Offloading Algorithm with Incremental Learning - 1

Input: *task*

Output: *Selected Offloading Destination*

```

1:  $SGD_{edge} \leftarrow$  new SGD()
2:  $SGD_{cRSU} \leftarrow$  new SGD()
3:  $SGD_{cCN} \leftarrow$  new SGD()
4:  $SGD_{edge}.buildModel(Batch_{edge})$ 
5:  $SGD_{cRSU}.buildModel(Batch_{cRSU})$ 
6:  $SGD_{cCN}.buildModel(Batch_{cCN})$ 
7: procedure SELECTOFFLOADINGDESTINATION(task)
8:    $t_{edge} \leftarrow SGD_{edge}.predict(Context_{edge})$ 
9:    $t_{cRSU} \leftarrow SGD_{cRSU}.predict(Context_{cRSU})$ 
10:   $t_{cCN} \leftarrow SGD_{cCN}.predict(Context_{cCN})$ 
11:  server  $\leftarrow$  server with minimum predicted  $t_{server}$ 
12:  task.setAssociatedServer(server)
13:  taskDictionary.put(task.id, Context_{server})
14:  offload task to server
15: end procedure
16: procedure TASKCOMPLETED(task)
17:  server  $\leftarrow$  task.getAssociatedServer()
18:  id  $\leftarrow$  task.getID()
19:   $t_o \leftarrow$  observed response time
20:  if server = Edge then
21:     $Context_{edge} = taskDictionary.remove(id)$ 
22:     $SGD_{edge}.update(Context_{edge}, t_o)$ 
23:  end if
24:  if server = cloudRSU then
25:     $Context_{cRSU} = taskDictionary.remove(id)$ 
26:     $SGD_{cRSU}.update(Context_{cRSU}, t_o)$ 
27:  end if
28:  if server = cloudCN then
29:     $Context_{cCN} = taskDictionary.remove(id)$ 
30:     $SGD_{cCN}.update(Context_{cCN}, t_o)$ 
31:  end if
32: end procedure
33: procedure TASKFAILED(task)
34:  server  $\leftarrow$  task.getAssociatedServer()
35:  id  $\leftarrow$  task.getID()
36:   $t_o \leftarrow$  observed response time
37:  if server = Edge then
38:     $Context_{edge} = taskDictionary.remove(id)$ 
39:     $SGD_{edge}.update(Context_{edge}, t_p)$ 
40:  end if
41:  if server = cloudRSU then
42:     $Context_{cRSU} = taskDictionary.remove(id)$ 
43:     $SGD_{cRSU}.update(Context_{cRSU}, t_p)$ 
44:  end if
45:  if server = cloudCN then
46:     $Context_{cCN} = taskDictionary.remove(id)$ 
47:     $SGD_{cCN}.update(Context_{cCN}, t_p)$ 
48:  end if
49: end procedure

```

cessfully returned from the selected server, the *taskCompleted* procedure (lines 16-32) will be called. This procedure will first retrieve the associated task's meta-data (i.e., the server chosen for offloading and taskID) (lines 17-18). It also makes use of the true observed response time (t_o) (line 19). Once the associated server is identified, the task dictionary will be

Algorithm 6 Task Offloading Algorithm with Incremental Learning - 2

Input: *task*

Output: Selected Offloading Destination

```
1: numApps  $\leftarrow$  Number of application types
2: for i  $\leftarrow$  0 to numApps - 1 do
3:   SGDedge[i]  $\leftarrow$  new SGD()
4:   SGDcRSU[i]  $\leftarrow$  new SGD()
5:   SGDcCN[i]  $\leftarrow$  new SGD()
6:   SGDedge[i].buildModel(Batchedge[i])
7:   SGDcRSU[i].buildModel(BatchcRSU[i])
8:   SGDcCN[i].buildModel(BatchcCN[i])
9: end for
10: procedure SELECTOFFLOADINGDESTINATION(task)
11:   type  $\leftarrow$  task.getApplicationType()
12:   tedge  $\leftarrow$  SGDedge[type].predict(Contextedge)
13:   tcRSU  $\leftarrow$  SGDcRSU[type].predict(ContextcRSU)
14:   tcCN  $\leftarrow$  SGDcCN[type].predict(ContextcCN)
15:   server  $\leftarrow$  server with minimum predicted tserver
16:   task.setAssociatedServer(server)
17:   taskDictionary.put(task.id, Contextserver)
18:   offload task to server
19: end procedure
20: procedure TASKCOMPLETED(task)
21:   server  $\leftarrow$  task.getAssociatedServer()
22:   id  $\leftarrow$  task.getID()
23:   to  $\leftarrow$  observed response time
24:   type  $\leftarrow$  task.getApplicationType()
25:   if server = Edge then
26:     Contextedge = taskDictionary.remove(id)
27:     SGDedge[type].update(Contextedge, to)
28:   end if
29:   if server = cloudRSU then
30:     ContextcRSU = taskDictionary.remove(id)
31:     SGDcRSU[type].update(ContextcRSU, to)
32:   end if
33:   if server = cloudCN then
34:     ContextcCN = taskDictionary.remove(id)
35:     SGDcCN[type].update(ContextcCN, to)
36:   end if
37: end procedure
38: procedure TASKFAILED(task)
39:   server  $\leftarrow$  task.getAssociatedServer()
40:   id  $\leftarrow$  task.getID()
41:   to  $\leftarrow$  observed response time
42:   type  $\leftarrow$  task.getApplicationType()
43:   if server = Edge then
44:     Contextedge = taskDictionary.remove(id)
45:     SGDedge[type].update(Contextedge, tp)
46:   end if
47:   if server = cloudRSU then
48:     ContextcRSU = taskDictionary.remove(id)
49:     SGDcRSU[type].update(ContextcRSU, tp)
50:   end if
51:   if server = cloudCN then
52:     ContextcCN = taskDictionary.remove(id)
53:     SGDcCN[type].update(ContextcCN, tp)
54:   end if
55: end procedure
```

accessed to obtain the context associated with the received task. Then, the associated model will be updated using SGD-based parameter update (equation 12) (lines 20-31). In order to maintain a relatively small size of the task dictionary, the entry associated with the returned task will be deleted from dictionary upon the receipt of that task's results. When the offloaded task fails, the *taskFailed* procedure (lines 33-49) is called. This procedure operates in a manner that resembles that of the *taskCompleted* procedure. However, the associated model is updated with a penalty value (t_p). This value is set such that the associated model is updated in a way that forces it to predict a high value of response time for upcoming offloading requests. Such a high predicted value would potentially prevent the offloading engine from choosing the respective server for subsequent tasks.

The other variant of the incremental learning-based algorithm (algorithm 6) maintains, for each server, an array of SGD-based models. Each model in the array can be used to make response time predictions for a particular application type. As shown, this algorithm initializes and fits preliminary models for different application types (lines 2-9). On the other hand, the three essential procedures in this algorithm are similar to those of algorithm 5; the only distinction is that these procedures will first identify the application type to which the task belongs and then use the associated model accordingly.

IV. RESULTS AND ANALYSIS

In order to assess the performance of the proposed algorithms, they have been implemented and evaluated using the EdgeCloudSim simulation tool [32]. EdgeCloudSim provides a simulation environment for Mobile Edge computing (MEC) and VEC systems [33], [6]. It allows modeling of computational servers, network infrastructure and mobile vehicles. It also allows users to defined different application types with varying characteristics. The vehicular mobility model assumed in this work is similar to that of [6]. In this model, its is assumed that a 16 km road is divided into 40 400-meter segments with each segment having a dynamic velocity value and covered by a single RSU. Hence, the speed of a vehicle dynamically varies based on the type of segment it moves on. This allows to differentiate the traffic density on each segment of the road and, consequently, the demand placed on the associated vehicular resources especially the edge servers (i.e., RSUs) covering different road segments and their associated network connection's bandwidth. When the simulation is started, vehicles are assigned random locations on the road and move in a single direction with a predefined segment-dependent speed. In addition, the road is defined as a circular route keeping the number of vehicles the same for the entire simulation time. This work assumes a VEC system with three representative application types, namely, traffic management, danger assessment and infotainment applications. Application characteristics are shown in Table II. Configuration parameters of the computational servers and network resources are shown in Table III.

The presented algorithms are compared to other existing algorithms that include the MAB-based algorithm (MAB) [12], game theory-based (Game-Theory) [15], machine learning-based (ML_based) besides the time series forecasting-based

TABLE II. VEHICULAR APPLICATION CHARACTERISTICS

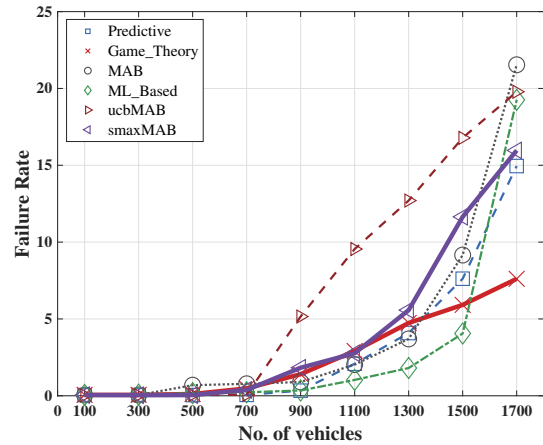
	Traffic Management	Danger Assessment	Infotainment Application
Vehicles percentage	30%	35%	35%
Input file size (KB)	20	40	20
Output file size (KB)	20	20	80
Inter-arrival time (second)	3	5	15
Instruction count ($\times 10^9$)	3	10	20
Utilization on Edge VM (%)	6	20	40
Utilization on Cloud VM (%)	1.6	4	8
Delay sensitivity (α)	0.6	0.90	0.35
Maximum delay requirement (second)	0.50	1.25	1.75
Penalty value (t_p)	1.25	2.25	2.5

TABLE III. SIMULATION PARAMETERS

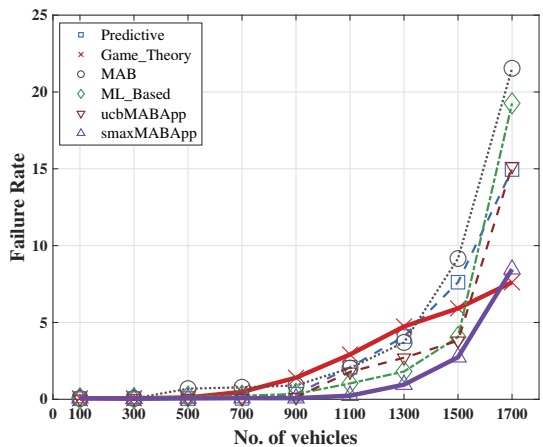
Parameter	Value
Simulation time (minutes)	60
Number of vehicles	100 - 1700
Vehicles counter size	200
No. of virtual Machines (VMs) on cloud server	20
No. of VMs on edge server	2
Processing capacity of cloud VM (MIPS)	75000
Processing capacity of edge VM (MIPS)	20000
WLAN range (meter)	200
WLAN bandwidth (Mbps)	100
MAN bandwidth (Mbps)	1000
WAN bandwidth (Mbps)	50
WAN propagation delay (second)	0.15
CN bandwidth (Mbps)	20
CN propagation delay (second)	0.16
Maximum reward (R_{max})	1
Pre-training batch size (instances)	100

(Predictive) algorithms [6]. Comparison results cover all possible implementations of algorithms 3 and 4. On the one hand, there are two possible implementations of algorithm 3; using either UCB or soft-max algorithms, denoted as *ucbMAB* and *smaxMAB*, respectively. On the other hand, algorithm 4 can also be implemented using either UCB (denoted as *ucbMABApp*) or soft-max (denoted as *smaxMABApp*). In addition, algorithm 5 is denoted as *SGDArm* while algorithm 6 is denoted as *SGDApp*. In VEC systems, failure rate is an important factor in assessing the performance of different task offloading schemes. Typically, an offloaded task would fail if a virtual machine (VM), on the selected server, has very high utilization that prevents it from executing the offloaded task, or if the available network bandwidth of the selected server is not sufficient to upload/download the input/output of the offloaded task. In other words, an offloaded task can fail due to unavailability of computational or networking resources. Fig. 4 shows and compares the average task failure rate, among all application types, under different offloading algorithms, as the number of vehicles is increased from 100 to 1700. In general, the failure rate increases as the number of vehicles is increased. As shown in Fig. 4a, the proposed MAB-based offloading algorithms with application-dependent reward (i.e., *ucbMAB* and *smaxMAB*) perform reasonably well - in terms of failure rate for small to moderate number of vehicles (≤ 700), as compared to other competitor algorithms. However, their associated failure rates increase significantly as the number of vehicles increases beyond 700 vehicles. On the other hand, Fig.

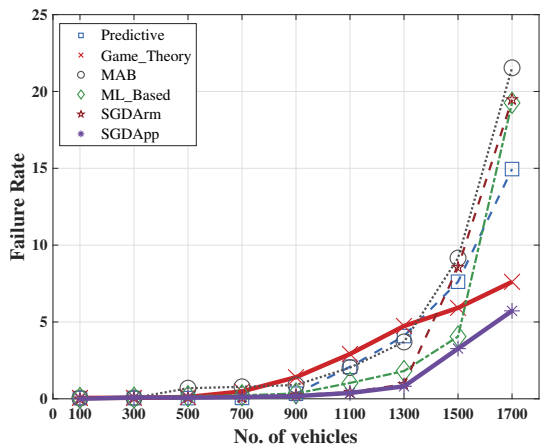
4b compares the failure rate of the MAB-based offloading with separate application-dependent MAB learners (i.e., algorithm 4) to that of other existing algorithms.



(a)



(b)



(c)

Fig. 4. Failure Rate Comparison: (a) Algorithm 3, (b) Algorithm 4, (c) Algorithms 5 and 6.

As shown, the *smaxMABApp* has outperformed the vast majority of other algorithms in terms of failure rate; having a dedicated MAB learner per application type allows the agent to devise an efficient offloading policy in which offloading decisions are based on the offloading history of similar tasks. While the *ucbMABApp* algorithm has achieved relatively low failure rate when the number of vehicles is less than 900, its performance has deteriorated in response to increasing the number of vehicles. Furthermore, Fig. 4c compares the failure rate under the proposed *SGDArm* and *SGDApp* algorithms to that under other algorithms. As shown, the *SGDArm* algorithm, in which an online SGD learner is maintained per each offloading option, has outperformed all competitor algorithms when the number of vehicles is less than 1300. However, its performance has significantly dropped beyond 1300 vehicles; as a single model is shared among all application types, updating the SGD-based model with a penalty value associated with a particular application type may adversely affect the offloading decision for subsequent tasks with different types. On the other hand, the *SGDApp* algorithm, in which an array of SGD-based learners is maintained, has shown significant improvement in failure rate especially under high number of vehicles (i.e., ≥ 1500). This can be due to that fact that *SGDApp* maintains an array of learners per each server with a dedicated model for each application type and, consequently, ensures that model updates are performed due to tasks with similar characteristics. Hence, the *SGDApp* algorithm has maintained consistent learning pattern achieving an efficient utilization of available contextual information and previous offloading history to steer current offloading decisions. For the other competitor algorithms, the failure rate of ML_Based and the Predictive algorithms [6], MAB [12] have significantly increased for high number of vehicles (i.e., ≥ 1500). On other hand, the game theory-based algorithm [15] has witnessed a noticeable linear increase in failure rate as the number of vehicles is increased beyond 900 vehicles. As compared to other algorithms especially in the more congested situation (i.e., no. of vehicles = 1700), *SGDApp* has achieved a failure rate reduction that ranges from 24.81%, as compared to GAME_Theory [15], to 73.43% as compared to MAB [12]. As for the other competitor algorithms, they tend to have acceptable failure rate values when the number of vehicles is less than 900. However, their associated failure rates start increasing after 900 vehicles. For instance, the ML_based algorithm offloads tasks with small instruction count to the nearby edge servers and the tasks with larger instruction count to the remote cloud server. However, the ML_based algorithm relies on statically trained models that do not gain any knowledge from the outcomes of the online offloading decisions. Hence, as the VEC system becomes more congested, its failure rate become worse. On the other hand, the MAB-based algorithm in [12] is also aware of the task's instruction count. However, it falls short when the VEC system becomes congested, and consequently, the failure rate starts increasing. In addition, the game theory-based algorithm tends to offload the majority of tasks to the edge servers regardless of their type. Hence, the lack of computing capacity at the edge servers lead to more task failure. Therefore, the task failure situation is noticeable after 900 vehicles and increases linearly, with respect to the number of vehicles. Furthermore, the predictive algorithm does not consider task characteristics when making offloading decisions; it increases the probability of selecting an

offloading destination that has recently provided better results. Hence, such a strategy would become inadequate when the VEC system becomes overloaded.

Admittedly, service time (a.k.a. response time) is another important evaluation metric. It represents the total time required to offload a task and obtain its results. Fig. 5 shows and compares the average service time values obtained under different offloading algorithms, with respect to the number of vehicles, for the successfully executed tasks.

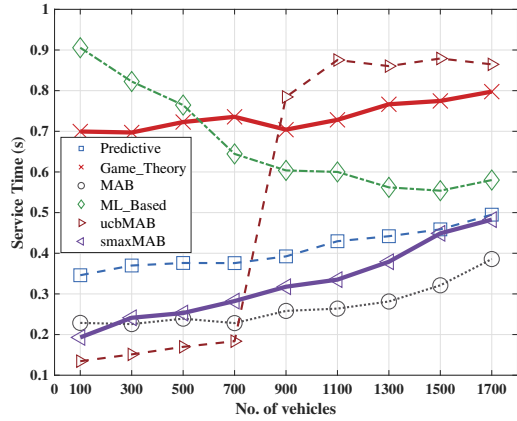
In general, the average service time increases with respect to the number of vehicles with the MAB algorithm [12] yielding the lowest service time among all competitor algorithms. As shown in Fig. 5a, the proposed *ucbMAB* algorithm has outperformed all other counterparts when the number of vehicles is less than 700 but its performance has dropped significantly after 700 vehicles. Similarly, the proposed *smaxMAB* algorithm has attained relatively acceptable service time values for systems in which the number of vehicles is less than 700. On the other hand, Fig. 5b illustrates that the proposed *ucbMABApp* and *smaxMABApp* algorithms have shown a profound ability to minimize average service time values for systems with less than 700 vehicles.

While the performance of *ucbMABApp* has dropped after 700 vehicles, *smaxMABApp* has consistently maintained a comparable performance to that of its MAB counterpart in [12]. Furthermore, Fig. 5c shows that the proposed *SGDArm* and *SGDApp* algorithms surpass all their counterparts for systems with at most 700 vehicles. However, as the VEC system becomes more congested, their obtained service time values increase. Nevertheless, the *SGDApp* algorithm has maintained a steadily comparable service time to that of the MAB algorithm in [12]. It is worth noting that although the MAB algorithm presented in [12] has achieved low average service time for the successfully executed tasks, it has suffered significant increases in failure rate in congested VEC systems as shown in Fig. 4.

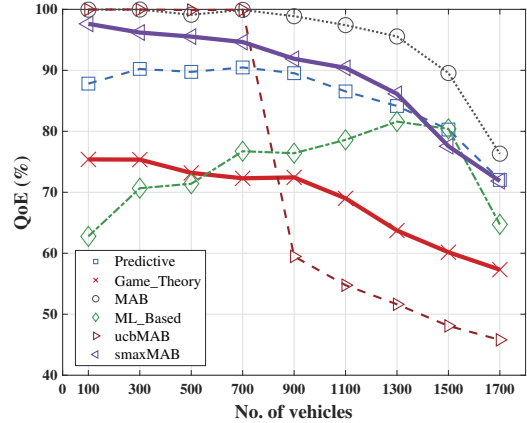
Although the failure rate and average service time provide adequate measures to evaluate task offloading algorithms, considering them as individual metrics may provide misleading results in systems where offloaded tasks may be lost. For instance, it may not be acceptable to have a low average service time for successfully executed tasks while having a high failure rate. Therefore, this work utilizes the Quality of Experience (QoE) formula proposed in [6], which combines both the service time and task failure as shown in equation 13.

$$QoE_i = \begin{cases} 0, & \text{if } i \text{ has failed} \\ 0, & \text{if } T_i \geq 2T_{max_i} \\ (1 - \frac{T_i - T_{max_i}}{T_{max_i}}) \cdot (1 - \alpha_A) \times 100\%, & \text{if } T_{max_i} \leq T_i < 2T_{max_i} \\ 100\%, & \text{if } T_i \leq T_{max_i} \end{cases} \quad (13)$$

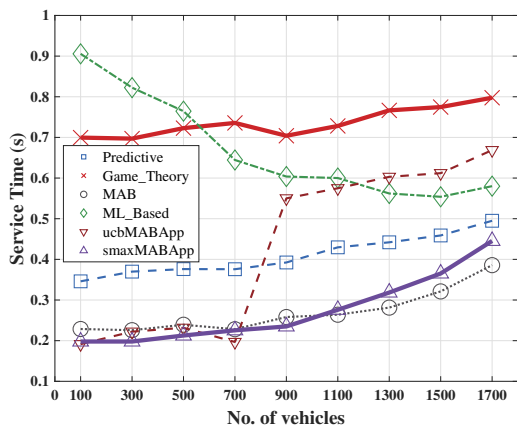
where T_i is the response time of task i , T_{max_i} is the maximum delay requirement of that task and α_A is that task's delay sensitivity. Evidently, the average QoE value decreases when task i is completed later than its associated delay requirement. If the observed service time exceeds twice the tasks' delay requirement or if task i fails, the QoE value is



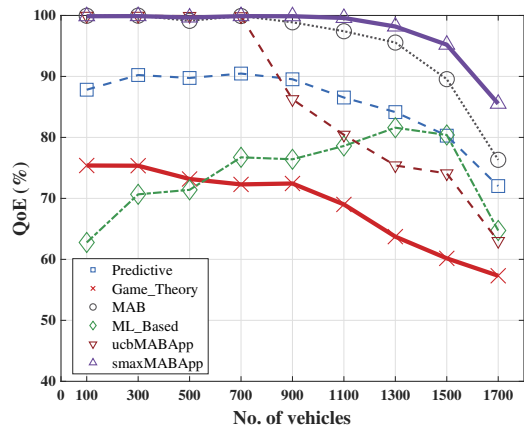
(a)



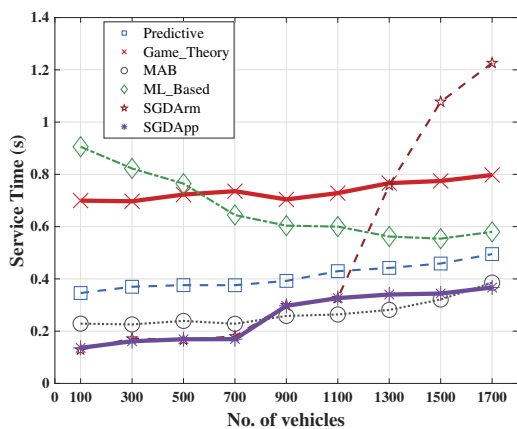
(a)



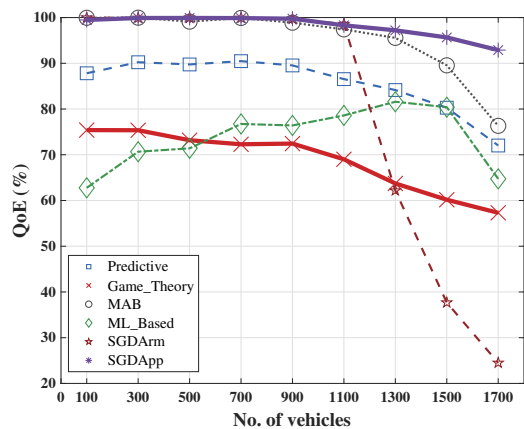
(b)



(b)



(c)



(c)

Fig. 5. Service Time Comparison: (a) Algorithm 3, (b) Algorithm 4, (c) Algorithms 5 and 6.

Fig. 6. Quality of Experience (QoE) Comparison: (a) Algorithm 3, (b) Algorithm 4, (c) Algorithms 5 and 6.

set to 0. Hence, the QoE metric provides a unified metric to match the observed service time to the delay requirements of different tasks besides assessing the balance between service time and failure rate. Fig. 6 depicts the average QoE values as a function of the number of vehicles.

As shown in Fig. 6a, the proposed *ucbMAB* and its MAB counterpart have provided the maximum QoE (100%) when the number of vehicles is low i.e., less than 700; because of their low response time and failure rate. Similarly, the proposed *ucbMAXApp* and *smaxMABApp* algorithms besides the

MAB algorithm in [12] have outperformed other algorithms in terms of QoE when the number of vehicles is less than 700, as shown in Fig. 6b. In addition, the *smaxMABApp* algorithm has maintained a reasonably higher QoE values for more congested systems i.e., when the number of vehicles exceeds 700. Furthermore, Fig. 6c demonstrates the ability of the proposed *SGDApp* and *SGDApp* algorithms to achieve a 100% QoE for systems with up to 1100 vehicles, with the *SGDApp* algorithm maintaining its superiority over other algorithms beyond 1100 vehicles. As shown in Fig. 6b and 6c, the proposed *smaxMABApp* and *SGDApp* algorithms have achieved 12.05% and 21.70% improvement in QoE as compared to their MAB counterpart, respectively, when the number of vehicles is equal to 1700.

Considering the competitor algorithms, the ML_based, predictive and game-theory-based algorithms provide the lowest QoE values. On the one hand, the ML_based algorithm is not able to respond to the dynamic changes of the VEC environment such as network bandwidth and server utilization. In other words, it is not able to dynamically adjust its offloading policy as it depends on statically trained models that would yield poor performance if the run-time conditions differ from those observed during offline model building. On the other hand, the game theory-based provides poor QoE readings because the main objective of the game model is neither to minimize the service time nor to improve failure rate but rather to attain a stable equilibrium. For the predictive algorithm, its QoE values never exceed 90% even with no task failure as it does not essentially minimize service time. Furthermore, The MAB-based algorithm (i.e., MAB [12]) does not guarantee the delay requirements of different task types as the number of vehicles exceeds 900. This can be due to the mismatch between the offloaded task's processing demand and the selected offloading destination violating that task's delay requirements.

Apparently, the failure rate, service time and QoE results of the proposed *SGDApp* algorithm prove the ability of the proposed contextual incremental learning scheme to handle task offloading in VEC systems with diversified application characteristics. In fact, incremental learning allows the task offloading algorithm to dynamically construct a robust offloading policy that efficiently utilizes the available contextual information and offloading history to guide subsequent offloading decisions. In other words, incremental learning allows the constructed models to gain new knowledge at run-time and vary model parameters in accordance with recently observed offloading outcomes. In order to prove the ability of the *SGDApp* algorithm to behave in a VEC system with different application type, its behaviour has further been analyzed and compared to its MAB counterpart [12], as the later has shown almost the best performance among other competitor algorithms. In this regard, Fig. 7 shows the task offloading distribution, considering all application types, under the *SGDApp* and the MAB algorithms. It shows the percentage of tasks offloaded to each of the edge and cloud servers. As shown, the proposed *SGDApp* algorithm has maintained more balanced utilization of edge and cloud units (Fig. 7a) as compared to its MAB counterpart (Fig. 7b); it has offloaded an almost identical proportion of tasks to each of the edge and cloud servers.

In order to gain further insight about the offloading behavior for different application types. Fig. 8 and 9 show the

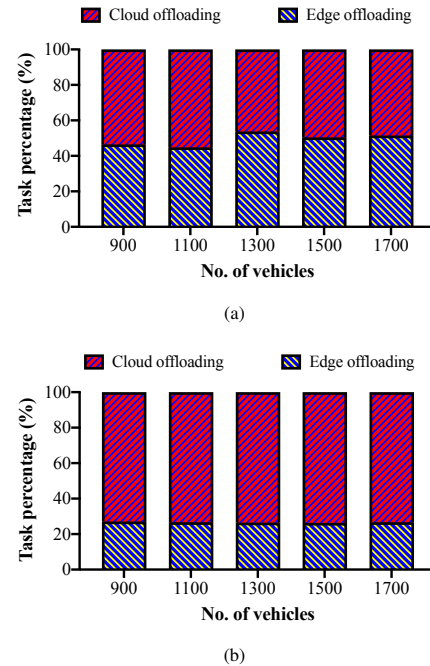
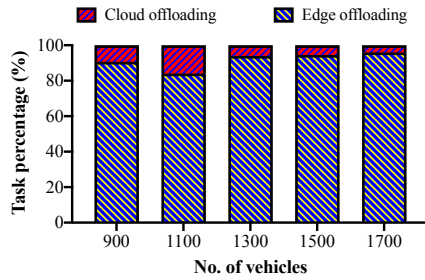


Fig. 7. Offloaded Tasks Distribution - all Application Types: (a) *SGDApp*, (b) MAB [12].

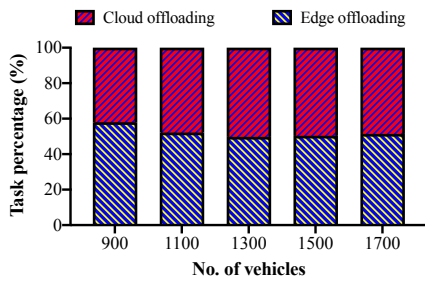
task offloading distributions for the traffic management and infotainment applications, respectively, under the *SGDApp* and the MAB algorithms. As shown in Table II, the traffic management application is characterized by having small tasks i.e., tasks with lower instruction count and average input/output file sizes. On the other hand, the infotainment application has larger tasks i.e., tasks with higher instruction count and average input/output file size, as compared to other application types.

As shown in Fig. 8a, the *SGDApp* algorithm tends to send the vast majority of the small tasks - generated from the traffic management application to the nearby edge servers. However, the MAB algorithm sends an almost equal proportion of the small tasks to each of the edge and cloud servers, as shown in Fig. 8b. On the other hand, Fig. 9a illustrates that the *SGDApp* algorithm sends the vast majority of the large infotainment tasks to the cloud server as opposed to MAB algorithm that sends all infotainment tasks to the cloud server, as shown in Fig. 9b.

Therefore, it can be observed that *SGDApp* is able to construct for each application type a model that would utilize the available contextual information to better steer that application's offloading decisions. In other words, the relatively small instruction count and file size of the traffic management tasks align with the processing and bandwidth capabilities of the edge servers. Consequently, sending these tasks to the edge servers has saved more cloud's processing capacity and network bandwidth for the processing- and bandwidth-hungry tasks such as infotainment tasks. On the other hand, sending small tasks to the cloud server in case of the MAB algorithm was harmful for all applications; it has caused higher service times for the small tasks and resulted in more resource contention with the large tasks on the cloud resources.

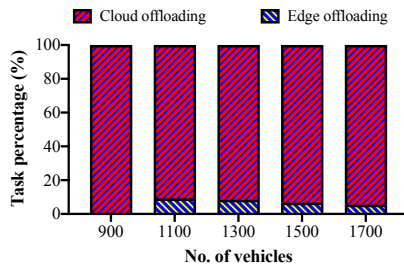


(a)

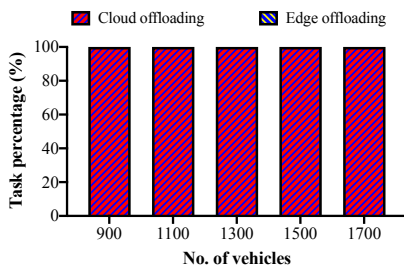


(b)

Fig. 8. Offloaded Tasks Distribution - Traffic Management: (a) SGDApp, (b) MAB [12].



(a)



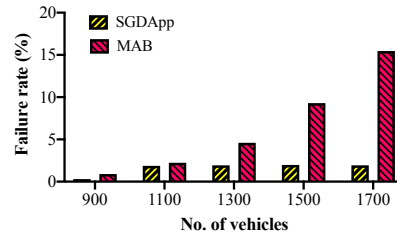
(b)

Fig. 9. Offloaded Tasks Distribution - Infotainment: (a) SGDApp, (b) MAB [12].

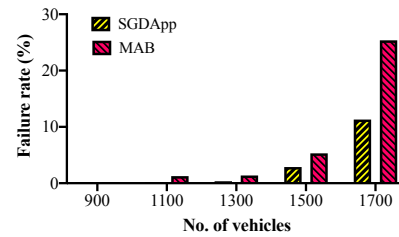
Ultimately, the offloading decision made for each application type has a direct consequence on that application's failure rate and QoE metrics. As shown in Fig. 10, the proposed *SGDApp* algorithm has obtained noticeable improvement in failure rate as compared to its counterpart. It has achieved better failure rate values for both the small traffic management tasks (Fig. 10a) and the large infotainment tasks (Fig. 10b).

The *SGDApp* algorithm has achieved up to 87.5% and 55.4% improvement in failure rate for the traffic management and infotainment applications, respectively, when the number of vehicles is 1700.

Similarly, the proper offloading decisions made by the *SGDApp* algorithm has achieved better QoE for both small and large tasks especially in more loaded VEC systems with the number of vehicles exceeding 1300, as shown in Fig. 11.

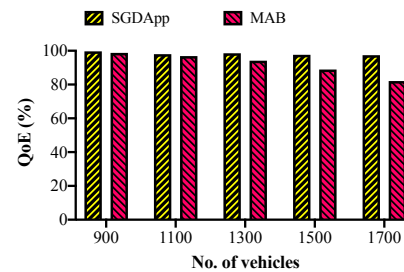


(a)

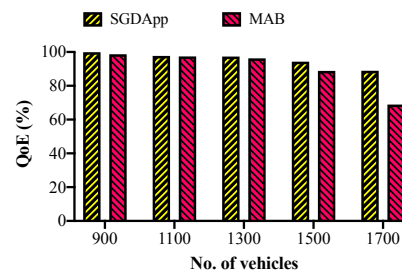


(b)

Fig. 10. Failure Rate of Different Applications: (a) Traffic Management, (b) Infotainment.



(a)



(b)

Fig. 11. QoE of Different Applications: (a) Traffic Management, (b) Infotainment.

The proposed *SGDApp* algorithm has consistently maintained higher QoE values for both application types, with respect to the number of vehicles. As depicted in Fig. 11a, the *SGDApp* algorithm has achieved up to 18.7% improvement in QoE, as compared to its counterpart, in a VEC system with 1700 vehicles. On the other hand, Fig. 11b illustrates the superiority of the *SGDApp* algorithm; its QoE improvement has reached 29.02% as the VEC system becomes more congested, with 1700 vehicles.

V. CONCLUSION

Vehicular Edge Computing (VEC) systems have recently been introduced to provide a seamless integrated computing platform to execute various kinds of vehicular applications. In these systems, computational tasks generated from in-vehicle applications are offloaded to either the edge or the cloud servers. In addition, VEC systems are characterized by a dynamically changing resource utilization besides having to handle diversified application types. Hence, an efficient task offloading scheme is required to ensure appropriate selection of offloading destinations, considering both application characteristics and the status of VEC resources. Therefore, this paper has presented a number of Multi-Armed Bandit (MAB) algorithms for task offloading in VEC systems with a representative set of applications. The rationale behind the proposed algorithms is to utilize contextual information such as application type and current resource utilization to achieve efficient application-specific offloading decisions. The proposed algorithms were implemented based on either a single MAB learner with application-dependent reward formulation, multiple dedicated MAB learners with a specific learner for each application type or a contextual bandits approach - based on incremental learning methods. The proposed algorithms were thoroughly analyzed and compared to other closely related task offloading algorithms. Simulation results proved the ability of the proposed contextual bandits-based algorithm to surpass all other algorithms under the failure rate and QoE metrics besides achieving adequately comparable service time values. In addition, it demonstrated the ability to efficiently utilize the available VEC resources and make the most appropriate decision for each application type, considering the interplay between application characteristics, timing requirements, server's computational capacity and network status. Hence, utilizing contextual information to dynamically construct and adjust incremental learning models has proved its feasible applicability for task offloading in VEC systems.

REFERENCES

- [1] D.-K. Choi, J.-H. Jung, H.-B. Nam, and S.-J. Koh, "Agent-based in-vehicle infotainment services in internet-of-things environments," *Electronics*, vol. 9, no. 8, 2020.
- [2] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–15, 2020.
- [3] L. Liu, C. Chen, Q. Pei, S. Maharjan, and Y. Zhang, "Vehicular edge computing and networking: A survey," *Mobile Networks and Applications*, Jul 2020.
- [4] S. Raza, W. Liu, M. Ahmed, M. R. Anwar, M. A. Mirza, Q. Sun, and S. Wang, "An efficient task offloading scheme in vehicular edge computing," *Journal of Cloud Computing*, vol. 9, no. 1, p. 28, Jun 2020.
- [5] Y. Wang, S. Wang, S. Zhang, and H. Cen, "An edge-assisted data distribution method for vehicular network services," *IEEE Access*, vol. 7, pp. 147 713–147 720, 2019.
- [6] C. Sonmez, C. Tunca, A. Ozgovde, and C. Ersoy, "Machine learning-based workload orchestrator for vehicular edge computing," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2020.
- [7] R. Zhang, P. Cheng, Z. Chen, S. Liu, Y. Li, and B. Vucetic, "Online learning enabled task offloading for vehicular edge computing," *IEEE Wireless Communications Letters*, vol. 9, no. 7, pp. 928–932, 2020.
- [8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.
- [9] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Machine Learning: ECML 2005*, J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 437–448.
- [10] Y.-L. He, X.-L. Zhang, W. Ao, and J. Z. Huang, "Determining the optimal temperature parameter for softmax function in reinforcement learning," *Applied Soft Computing*, vol. 70, pp. 80–85, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494618302758>
- [11] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2x access technologies: Regulation, research, and remaining challenges," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1858–1877, 2018.
- [12] Y. Sun, X. Guo, J. Song, S. Zhou, Z. Jiang, X. Liu, and Z. Niu, "Adaptive learning-based task offloading for vehicular edge computing systems," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 3061–3074, 04 2019.
- [13] X. Xu, Y. Xue, X. Li, L. Qi, and S. Wan, "A computation offloading method for edge computing with vehicle-to-everything," *IEEE Access*, vol. 7, pp. 131 068–131 077, 2019.
- [14] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint load balancing and offloading in vehicular edge computing and networks," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4377–4387, 2019.
- [15] Y. Wang, P. Lang, D. Tian, J. Zhou, X. Duan, Y. Cao, and D. Zhao, "A game-based computation offloading method in vehicular multiaccess edge computing networks," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4987–4996, 2020.
- [16] P. Liu, J. Li, and Z. Sun, "Matching-based task offloading for vehicular edge computing," *IEEE Access*, vol. 7, pp. 27 628–27 640, 2019.
- [17] J. Feng, Z. Liu, C. Wu, and Y. Ji, "Mobile edge computing for the internet of vehicles: Offloading framework and job scheduling," *IEEE Vehicular Technology Magazine*, vol. 14, no. 1, pp. 28–36, 2019.
- [18] Z. Jiang, S. Zhou, X. Guo, and Z. Niu, "Task replication for deadline-constrained vehicular cloud computing: Optimal policy, performance analysis, and implications on road traffic," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 93–107, 2018.
- [19] Q. Hu, C. Wu, X. Zhao, X. Chen, Y. Ji, and T. Yoshinaga, "Vehicular multi-access edge computing with licensed sub-6 ghz, ieee 802.11p and mmwave," *IEEE Access*, vol. 6, pp. 1995–2004, 2018.
- [20] H. Peng and X. Shen, "Deep reinforcement learning based resource management for multi-access edge computing in vehicular networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2416–2428, 2020.
- [21] H. Sami, A. Mourad, and W. El-Hajj, "Vehicular-obus-as-on-demand-fogs: Resource and context aware deployment of containerized micro-services," *IEEE/ACM Transactions on Networking*, vol. 28, no. 02, pp. 778–790, mar 2020.
- [22] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 661–670.
- [23] C. Zhang, H. Wang, S. Yang, and Y. Gao, "A contextual bandit approach to personalized online recommendation via sparse interactions," in *Advances in Knowledge Discovery and Data Mining*, Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, and S.-J. Huang, Eds. Cham: Springer International Publishing, 2019, pp. 394–406.

- [24] A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang, "Contextual bandits with continuous actions: Smoothing, zooming, and adapting," in *Proceedings of the Thirty-Second Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, A. Beygelzimer and D. Hsu, Eds., vol. 99. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 2025–2027.
- [25] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.
- [26] C. Szepesvári, "Algorithms for reinforcement learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 4, no. 1, pp. 1–103, 2010.
- [27] L. Tang, Y. Jiang, L. Li, and T. Li, "Ensemble contextual bandits for personalized recommendation," in *Proceedings of the 8th ACM Conference on Recommender Systems*, ser. RecSys '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 73–80.
- [28] R. Cañamares, M. Redondo, and P. Castells, "Multi-armed recommender system bandit ensembles," in *Proceedings of the 13th ACM Conference on Recommender Systems*, ser. RecSys '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 432–436.
- [29] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Linear Regression*. New York, NY: Springer New York, 2013, pp. 59–126.
- [30] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Y. Lechevallier and G. Saporta, Eds. Heidelberg: Physica-Verlag HD, 2010, pp. 177–186.
- [31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10–18, Nov. 2009.
- [32] C. Sonmez, A. Ozgovde, and C. Ersoy, "Edgecloudsim: An environment for performance evaluation of edge computing systems," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 11, p. e3493, 2018.
- [33] C. Sonmez, A. Ozgovde, and C. Ersoy, "Fuzzy workload orchestration for edge computing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 769–782, 2019.

Body Weight Estimation using 2D Body Image

Rohan Soneja¹, Prashanth S², R Aarthi³
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

Abstract—Two dimensional images of a person implicitly contain several useful biometric information such as gender, iris colour, weight, etc. Among them, body weight is a useful metric for a number of usecases such as forensics, fitness and health analysis, airport dynamic luggage allowance, etc. Most current solutions for body weight estimation from images make use of additional apparatus like depth sensors and thermal cameras along with predefined features such as gender and height which generally make them more computationally intensive. Motivated by the need to provide a time and cost efficient solution, a novel computer-vision based method for body weight estimation using only 2D images of people is proposed. Considering the anthropometric features from the two most common types of images, facial and full body, facial landmark measurements and body joint measurements are used in deep learning and XG boost regression models to estimate the person's body weight. The results obtained, though comparable to previous approaches, perform much faster due to the reduced complexities of the proposed models, with facial models performing better than full body models.

Keywords—Body weight estimation; deep learning; xgboost regressor; anthropometric features; computer vision

I. INTRODUCTION

The purpose of this work is to estimate the weight of a person given only a two dimensional image. Many use cases necessitate the estimation of body weight without the physical measurement or presence of a person directly. For example, in health analysis to check the weight through mobile devices for a quick estimation, in forensics to gain additional identification features, in airports to estimate weight to aid dynamic baggage allowance, for physicians working remotely for rural patients, etc. Additionally, in social networks containing advertisements, considering the huge volume of images of people uploaded on the Internet daily, body weight can be taken as another ad-metric.

A novel cost-effective method that is computationally less intensive is proposed for estimating body weight using only the 2D images of a person and features extracted from the images. The person's image has the facial and the full body components of the person. Different procedures are applied to each type of image to obtain the weight of the person in the corresponding image.

For the model using only facial image a novel dataset, IDOC-Mugshots [1], containing over 70,000 frontal face images of prison inmates was used. This dataset, available on Kaggle, was obtained from the Illinois Department of Correction. In addition to that, VIP-Attribute dataset [2] containing 1000 images of celebrities was scraped from popular websites. For the model using full-body images Visual-Body-to-BMI dataset [3] containing around 6000 images scraped from the subreddit r/ProgressPics was used. A deep learning model and

an XGBoost regressor model were trained on the features extracted from both types of images and the results obtained showed the feasibility of using only 2D images for body weight estimation. The results obtained were analysed when using only facial images or full body images. They were comparable to previous work even with the reduced feature set and computationally less intensive methods used.

The remaining sections in the paper are organised as follows: first, we review related work done in the area and their results; this is followed by a brief description of the proposed architecture; then the implementation using features from face and full body is explained; finally a summary of the results and the conclusions drawn along with scope for future work are provided.

II. RELATED WORK

Previous approaches for this task either used additional apparatus such as RGB-D sensors or thermal cameras in addition to a camera to obtain more features than just the 2D image. In other cases, other features are used as input to the learning models that aren't obtained from the image such as gender, age, etc. These methods require prior data collection about the person or additional equipment and hence are neither fast nor cost efficient.

Some work has analyzed body weight or BMI from face images [4] [5] [6]. Wen et al. [7] first proposed a computational method for BMI prediction from face images based on the MORPH-II dataset, which obtained mean absolute errors for BMI in the range of 2.65-4.29 for different categories based on ethnicity. They also analyzed the correlations between facial features and BMI values. An Active Shape Model is used to extract facial features which are used to predict BMI using various regression techniques. Barr et al. [8] used facial landmarking to figure out adiposity (facial fatness) which positively correlates to the weight of the person. This method is less accurate in extreme underweight and obese cases though. A support vector machine regression model was used. Windhager et al. [9] showed that shape of the face has direct correlation with several body characteristics such as height and weight and can be determined by facial landmarking and spatial scaling. A total of 71 landmarks and semi landmarks were digitized to capture facial shape. Regression and geometric morphometric toolkit tool were used to estimate facial fatness. Additionally as a feature set, height measurement using anthropometer and saliva sample testing done apart from facial front photograph. Tai et al. [10] used Kinect sensors to estimate BMI using facial data on a regression model. Recently, Haritosh et al. [11] used convolution neural networks and artificial neural networks to estimate the weight using the facial image extracted using Viola-Jones detector.

There are a few studies on estimating human body weight or BMI from body related data [12] [13] [14], such as body measurements, 3-dimensional (3D) body data and RGB-D body images. Jiang et al. [3] used images of entire front body by scraping data from a Reddit page called r/ProgressPics. To estimate BMI they used anthropometric measurements from body contour segments with the help of skeletal joints and estimated the weight based on those features.

The body weight was studied directly by Velardo et al. [15] from anthropometric data collected by National Health And Nutrition Examination Survey, Centers for Disease Control and Prevention using a polynomial regression model to estimate the weight within 4% error using 2D and 3D data extracted from a low-cost Kinect RGB-D camera output. Pfitzner et al. [16] also used RGB-D camera data; this demonstrated a body weight estimation by volume extraction from RGB-D data with an accuracy of 79% for a cumulative error of $\pm 10\%$. Compared to a physician's estimation, this approach is already more suitable for drug dosing. Pichler et al. [17] estimated human body volume in clinical environment by eight stereo cameras around a stretcher and bioelectrical impedance analysis. Nguyen et al. [18] estimated body weight using a side view feature and a support vector regression model to obtain an average error of 4.62 kg for females and 5.59 kg for males.

III. PROPOSED ARCHITECTURE

Most of the previous research has been to estimate BMI using facial images and in some cases body weight. While that is shown to have a high correlation to the BMI and body weight, it is prone to high error. Without the use of external hardware, two methods to measure the body weight using a limited feature set are proposed. One is using face data, and the other is an extension of that which uses full body image data (including facial features).

Previous studies have shown that estimation of a person's weight given only the image of the person's face is possible due to the correlation of the weight with facial fatness, i.e. adiposity [19]. This can be measured by taking various measurements across the face such as the height and width of the face, length and width of the nose, etc.

Next, given the full body image of a subject, a novel method is chosen to extract features from the image. Instead of making anthropometric measurements with respect to body contours, measurements are made with respect to the bone joints of the person which are found to be correlated to the body weight [20]. Using these measurements result in several advantages.

- With current models, bone joint coordinates are easier to extract from images instead of the body contour, i.e., less computationally intensive.
- Bone joint coordinates are usually more accurate than body contour segments, i.e., more reliable measurements.
- Even in case of reliable body contour measurements, unlike that, bone joint coordinates are not affected by baggy/tight garments, i.e., less clothing bias.

These measurements are used in conjunction with the facial measurements as input features with the weight of the person being the target feature.



Fig. 1. Face Landmarks.

Finally, in both cases, a split of the feature set is used to train the machine learning model while the remaining is used to evaluate it, using the mean absolute error as the evaluation metric.

IV. IMPLEMENTATION

A. Features from Face

Weight estimation with only a 2D facial image was done by extracting measurements of various noticeable regions of the face such as eyes, eyebrows, nose, mouth and jawline. This was applied to two datasets, VIP-Attribute dataset [2] and a novel IDOC-Mugshots dataset [1]. Due to the varied nature of images and issues caused by background noise, first, face localisation is performed using the face detector in Python's dlib library. The cropped-out region of interest is then used for landmarking to detect key facial structures and thereby their measurements. For this task, the facial landmark detector included in the dlib library is used, which is internally implemented using an ensemble of regression trees [21]. This estimates 68 coordinates (x,y) on the face (Fig. 1) which are used to measure facial features such as:

- Left eyebrow width (18-22)
- Right eyebrow width (23 - 27)
- Left eye width (37 - 40)
- Right eye width (43 - 46)
- Nose width (32 - 36)
- Nose length (28 - 34)
- Outer lip width (49 - 55)
- Inner lip width (61 - 65)
- Face height (28 - 9)
- Face width (1 - 17)

Since the measurements are made from a 2-dimensional image there is a lack of perception of depth causing those

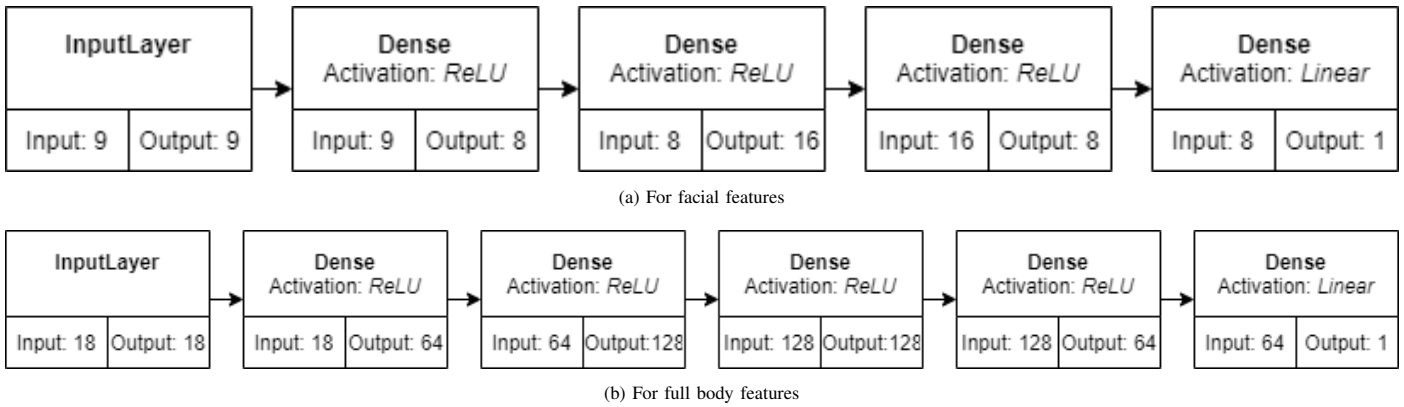


Fig. 2. Deep Learning Model Architectures.

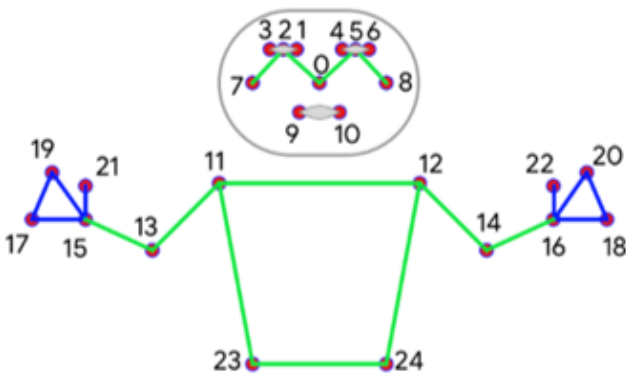


Fig. 3. Body Landmarks.

farther away in the image to have smaller measurements which hampers the performance of the model. Hence all the measurements were converted into ratios by dividing the lengths with the width of the face.

These extracted ratios exhibiting facial fatness (adiposity) are used as features for training. Hence no additional features other than the image itself is used for weight estimation. A deep learning model (Fig. 2a) and an XGBoost Regressor (40 estimators) were used on both datasets to estimate the body weight.

B. Features from Full Body

Initially, the BlazePose model [22] is used to detect important landmarks (Fig. 3) of the most prominent human body in the image that is provided. Since the dataset contains images with only one subject, the body joints detected are for the person of interest only. The dataset contains a lot of images that do not have the joints visible clearly. Based on the additional visibility parameter provided by MediaPipe’s Pose API, images that do not cross a certain threshold for all the joints that are considered are eliminated from training. There are four pairs of joints that are used, viz. shoulder, hip, elbow, and wrist. Various Euclidean distances are measured between the joints and they are scaled down with respect to the inter-shoulder distance (11 - 12) to ensure uniformity of measurements

between images of various resolutions and subject to image ratios.

- Left-shoulder to left-hip (11 - 23)
- Right-shoulder to right-hip (12 - 24)
- Left-hip to right-hip (23 - 24)
- Left-shoulder to right-hip (11 - 24)
- Right-shoulder to left-hip (12 - 23)
- Left-shoulder to left-elbow (11 - 13)
- Left-elbow to left-wrist (13 - 15)
- Right-shoulder to right-elbow (12 - 14)
- Right-elbow to right-wrist (14 - 16)

Finally, the features used for training are the aforementioned ratios along with the facial measurements as described in the facial model. It is important to note that the facial measurements taken on full body images are not accurate due to the fact that the face of the person takes up a very small region of the image as compared to the rest of the body.

A deep learning model (Fig. 2b) and an XGBoost Regressor (7 estimators) were used on the dataset to estimate the body weight.

V. RESULTS

The metric chosen to evaluate the model performance is Mean Absolute Error (MAE). MAE is calculated by taking the mean of absolute value of errors between the predicted weight and the actual weight of the person.

$$MAE(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Table I shows the MAE for each of the models and datasets. Overall and in case of facial features, VIP-Attribute dataset has yielded the best results with the deep learning model with $MAE = 9.8kg$ and in case of full body features, XGBoost performs slightly better than the deep learning model with $MAE = 18.2kg$. Fig. 4 shows the distribution of errors made by the various models for each of the datasets.

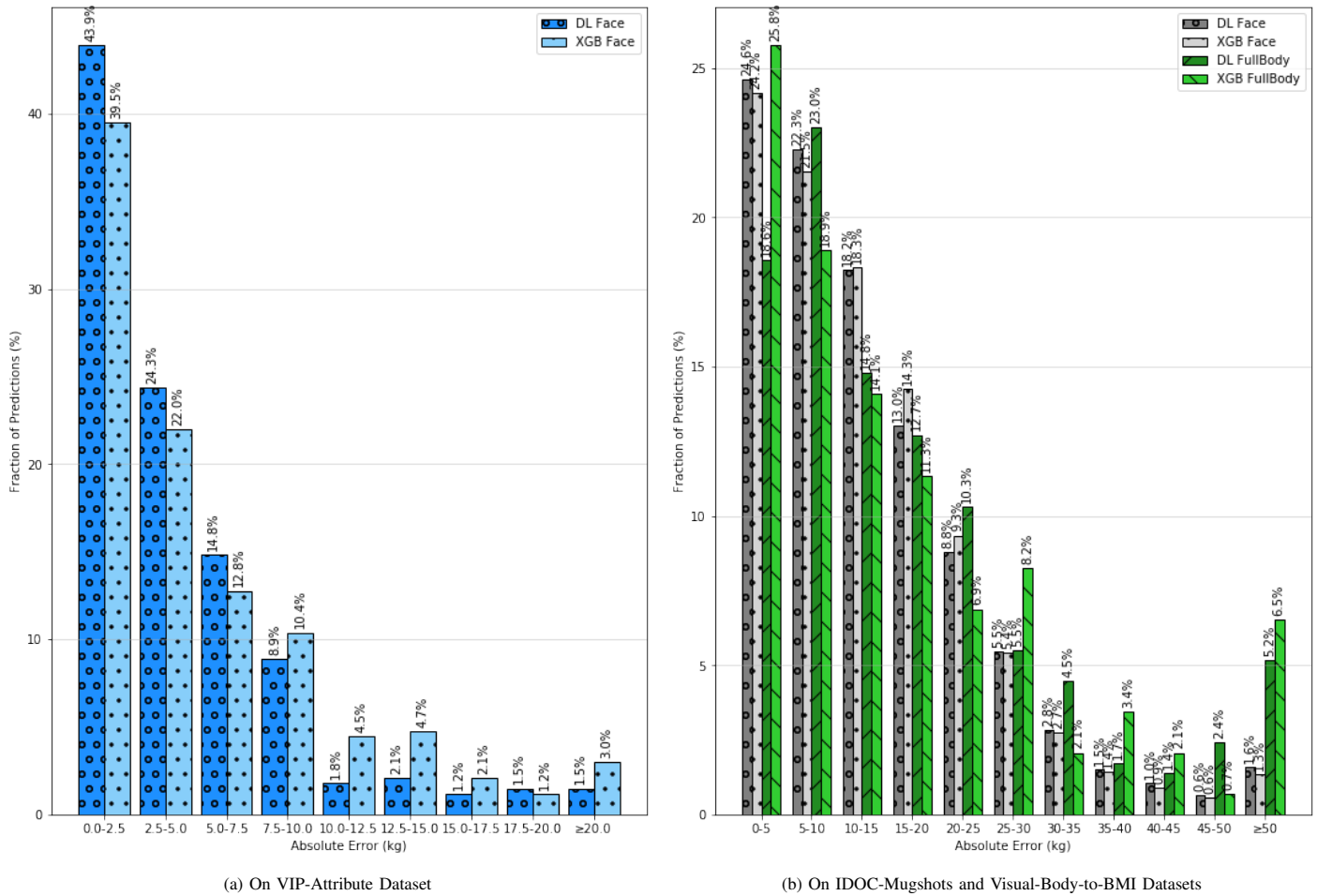


Fig. 4. Distribution of Errors for Weight Estimation.

TABLE I. MEAN ABSOLUTE ERRORS

S. N.	Dataset (only face)	Model	MAE (kg)
1.	VIP-Attribute	Deep Learning	9.8
2.	VIP-Attribute	XGBoost Regressor	11.9
3.	IDOC Mugshots	Deep Learning	13.5
4.	IDOC Mugshots	XGBoost Regressor	13.5

(A) MAE FOR FACE DATASET

S. N.	Dataset (full body)	Model	MAE (kg)
1.	Visual Body to BMI	Deep Learning	18.6
2.	Visual Body to BMI	XGBoost Regressor	18.2

(B) MAE FOR BODY DATASET

If the graph is skewed to the left, it indicates a better model performance since the errors are small in that case. The relative spike on the right end can be explained by the fact that the last bars represent errors more than the specified value and not just within a particular range. Fig. 5 shows the comparison between the ground truth body weights with their corresponding estimations as provided by the ML models in

the form of a scatter plot. The dotted line represents the ideal model with no error and the surrounding solid lines represent an absolute error of 15 kg. It can be seen that the models tend to underestimate weights more than 100 kg and slightly underestimate weights less than 60 kg.

Although it may seem unintuitive that the facial model performs better than the full body model (about 40% of predictions having error less than 2.5 kg (Fig. 4a) as opposed to around 25% of predictions having error less than 5 kg (Fig. 4b)), it is justified as follows.

- The facial datasets are much cleaner and regular due to faces being in the exact same positions for all images.
- Due to extensive occlusion in Visual Body to BMI owing to varying postures, object obstructions, etc., a lot of the measurements are approximated.
- The VIP-Attribute dataset has lesser range of body weights compared to the other datasets (Fig. 5).
- The full body feature set considers only bone joints which are inherently less correlated to the body weight as opposed to the body contour.
- The version of the Pose API that was used did not include bone joints below the waist, hence reducing

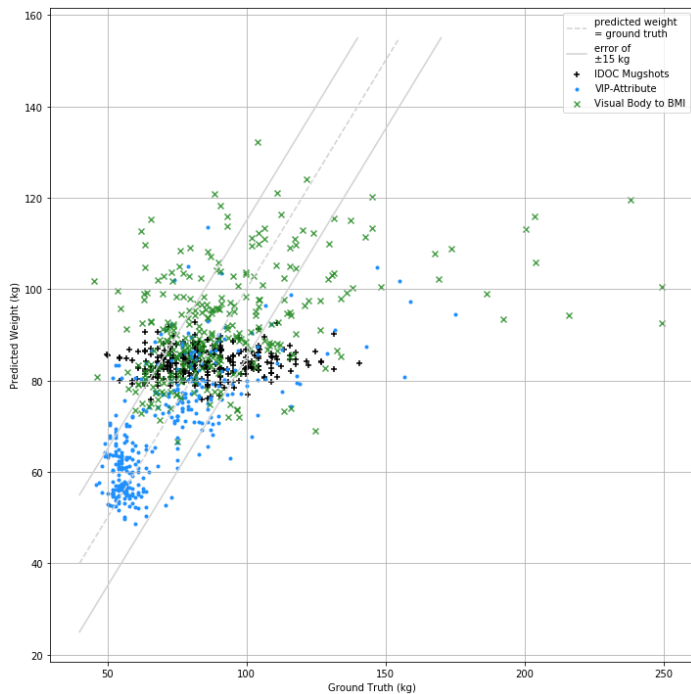


Fig. 5. Ground Truth v. Predicted Weight Scatterplot for Deep Learning Models.

the availability of femur length in the feature set, from which the models may have benefitted.

- The Visual Body to BMI dataset contains several instances of weights lying on the extreme ends (Fig. 5) due to the nature of the source of the dataset and hence in some cases, the error is too high (even more than 100 kg as evident from Fig. 4b).

With respect to the computational intensity, from experimentally timing the runtime of the feature extraction methods, it was found that extracting landmarks (using BlazePose [22]) instead of body contour (using CRF-as-RNN [23]) was 40 times faster, given the same hardware and running conditions. Of course, once the body contour is found, it requires more preprocessing (in the form of contour smoothing, pixel counting, length extraction, etc.), making it even more time consuming. Also, considering our model's reduced complexity using simpler features, we report a marginally higher MAE compared to 8.51 kg as reported by Dantcheva et al. [2] for the same VIP-Attribute dataset. Haritosh et al. [11] report an even higher MAE of 13.29 kg owing to the separate Reddit-HWBMI facial dataset that was used.

VI. CONCLUSION

In this work, the body weight of a person was estimated given just the image of the subject. Two types of datasets are employed viz. facial images and full body images. Features are extracted using publicly available libraries and they are used to train deep learning and XGB regressor models. The results obtained were compared and it was found that it is a viable and efficient method to estimate the body weight using just the person's image as long as the weights are not on

the higher extreme. To counter that problem, in the future, an efficient body contour detector may be developed that uses the landmarks to facilitate itself so that the model is not made computationally intensive. Once the Pose API is updated to support depth attributes to the joint coordinates, that too can be used as a feature among other currently unexplored features to improve the overall MAE.

REFERENCES

- [1] Elliot, "Idoc-mugshots dataset," 06 2018. [Online]. Available: <https://www.kaggle.com/elliottp/idoc-mugshots>
- [2] A. Dantcheva, F. Bremond, and P. Bilinski, "Show me your face and i will tell you your height, weight and body mass index," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 8 2018, pp. 3555–3560.
- [3] M. Jiang and G. Guo, "Body weight analysis from human body images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2676–2688, 10 2019.
- [4] K. Vikram and S. Padmavathi, "Facial parts detection using viola jones algorithm," in *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2017, pp. 1–4.
- [5] T. Keshari and S. Palaniswamy, "Emotion recognition using feature-level fusion of facial expressions and body gestures," in *2019 International Conference on Communication and Electronics Systems (ICCES)*, 2019, pp. 1184–1189.
- [6] N. Parameswaran and D. Venkataraman, "A computer vision based image processing system for depression detection among students for counseling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, pp. 503–512, 04 2019.
- [7] L. Wen and G. Guo, "A computational approach to body mass index prediction from face images," *Image and Vision Computing*, vol. 31, no. 5, pp. 392 – 400, 2013.
- [8] M. Barr, G. Guo, S. Colby, and M. Olfert, "Detecting body mass index from a facial photograph in lifestyle intervention," *Technologies*, vol. 6, no. 3, p. 83, 8 2018.
- [9] S. Windhager, F. L. Bookstein, E. Millesi, B. Wallner, and K. Schaefer, "Patterns of correlation of facial shape with physiological measurements are more integrated than patterns of correlation with ratings," *Scientific Reports*, vol. 7, no. 1, p. 45340, 5 2017.
- [10] C. Tai and D. Lin, "A framework for healthcare everywhere: Bmi prediction using kinect and data mining techniques on mobiles," in *2015 16th IEEE International Conference on Mobile Data Management*, vol. 2, 6 2015, pp. 126–129.
- [11] A. Haritosh, A. Gupta, E. S. Chahal, A. Misra, and S. Chandra, "A novel method to estimate height, weight and body mass index from face images," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 8 2019, pp. 1–6.
- [12] K. Padmavathi and S. Nithin, "Comparison of image processing techniques for detecting human presence in an image," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 383–388.
- [13] K. Hena, J. Amudha, and R. Aarthi, "A dynamic object detection in real-world scenarios," in *Proceedings of International Conference on Computational Intelligence and Data Engineering*, N. Chaki, N. Devarakonda, A. Sarkar, and N. C. Debnath, Eds. Singapore: Springer Singapore, 2019, pp. 231–240.
- [14] S. T. and P. B. Sivakumar, "Human gait recognition and classification using time series shapelets," in *2012 International Conference on Advances in Computing and Communications*, 2012, pp. 31–34.
- [15] C. Velardo and J.-L. Dugelay, "Weight estimation from visual body appearance," 10 2010, pp. 1 – 6.
- [16] C. Pfitzner, S. May, and A. Nüchter, "Evaluation of features from rgb-d data for human body weight estimation," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 10148 – 10153, 2017, 20th IFAC World Congress.
- [17] K. Santner, M. Rütger, H. Bischof, F. Skrabal, and G. Pichler, "Human body volume estimation in a clinical environment," 01 2009.

- [18] T. V. Nguyen, J. Feng, and S. Yan, "Seeing human weight from a single rgb-d image," *Journal of Computer Science and Technology*, vol. 29, no. 5, pp. 777–784, 9 2014.
- [19] L. Wen, G. Guo, and X. Li, "A study on the influence of body weight changes on face recognition," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–6.
- [20] C. Velardo, "Anthropometry and soft biometrics for smart monitoring," 2012, pp. 34–46.
- [21] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.
- [22] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.
- [23] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015. [Online]. Available: <http://dx.doi.org/10.1109/ICCV.2015.179>

A Detailed Study on the Choice of Hyperparameters for Transfer Learning in Covid-19 Image Datasets using Bayesian Optimization

Miguel Miranda¹, Kid Valeriano², José Sulla-Torres³
Universidad Nacional de San Agustín de Arequipa, Perú

Abstract—For many years, the area of health care has evolved, mainly using medical images to detect and evaluate diseases. Nowadays, the world is going through a pandemic due to COVID-19, causing a severe effect on the health system and the global economy. Researchers, both in health and in different areas, are focused on improving and providing various alternatives for rapid and more effective detection of this disease. The main objective of this study is to automatically explore as many configurations as possible to recommend a smaller starting hyperparameter space. Because the manual selection of these hyperparameters can lose configurations that generate more efficient models, for this, we present the MKCovid-19 workflow, which uses chest x-ray images of patients with COVID-19. We use knowledge transfer based on convolutional neural networks and Bayes optimization. A detailed study was conducted with different amounts of training data. This automatic selection of hyperparameters allowed us to find a robust model with an accuracy of 98% in test data.

Keywords—Transfer Learning; COVID-19; X-ray image; deep learning; Bayes optimization; machine learning; hyperparameter optimization

I. INTRODUCTION

Currently, there are different areas of knowledge where algorithms, heuristics, and artificial intelligence models are applied. The area of health care has evolved year after year, adopting the use of technology to save lives and improve life quality. Mainly visual information is frequently applied for the detection and evaluation of diseases. For that reason, the established fields of computer vision and medical imaging provide essential tools. The integration of these technologies and data analysis from various sources, real-time processing are core competencies necessary for the successful improvement of healthcare systems.

As we know, at the end of 2019, the first official cases of COVID-19 began to be reported in China, which continues to wreak havoc around the world today, mainly in the health system and the global economy. Currently, around 113 million cases of coronavirus (SARS-CoV-2) have been registered globally, 87 million cases of patients cured of COVID-19 and with a mortality of more than 2 million people¹. This virus is an epidemic disease, which can cause respiratory infections from a cold to serious illnesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS).

In order to better control the problems caused by COVID-19 and help reduce the death rate, the detection of this disease through medical imaging is an important factor. Currently, Chest X-rays (CXR) and Computed Tomography (CT) are commonly used, which allow the severity of the disease in the patient to be assessed and monitored.

At the beginning of the pandemic, at least in Latin American countries, such as Peru, serological and molecular tests were in short supply. They were caused by the difficulty of acquiring their central governments due to the high demand for purchase by the other countries. In these countries, due to the paucity of tests, chest radiographs were used more frequently; these were used because they are more accessible to patients. However, this presents a challenge for radiologists since pneumonia can be caused by other viruses or bacteria, making it difficult to diagnose and predict Covid-19 in the patient.

Nowadays, to face this challenge, computer-aided diagnostics (CAD) is used, accelerating and improving medical diagnosis precision. As part of this problem's solution, artificial intelligence algorithms are used due to the large-scale data processing capacity integrated into CAD. For this case, an analysis of medical images and deep learning is carried out. Specifically, transfer learning is used with models based on convolutional neural networks (CNN). Some of these models already used for related studies are RESNET50[1], COVID-NET[2], ResNet50V2 [3], VGG16 [4].

At the beginning of the pandemic, training data was not available, this being the limitation even though the authors used data augmentation techniques. In this work, data are collected from various sources to make an extensive data set for further study. Besides, Bayes optimization is used to carry out more in-depth research, such as the effect of using different amounts of data, the batch size, and other adjustments. Finally, it is studied and recommended that pre-trained models better adhere to the use of medical images to detect covid. This whole study aims to help choose the most optimal values of the hyperparameters that will directly influence the performance of these models.

All the above can help improve performance and speed to detect cases of Covid-19, as well as other diseases that can affect the lungs. This can be done by finding the best model for this type of "problem". For this study, Transfer Learning is used, explained, and detailed in this article's content, where the results obtained are shown.

After this introduction, this article is organized as follows: The works related to this article are explained in Section II,

¹<https://www.statista.com/statistics/1087466/covid19-cases-recoveries-deaths-worldwide>

the methods and material used in this research in Section III, the methodology in Section IV, and the experimental setups in Section V, experimental result in Section VI, conclusions in Section VII, and finally, future works in Section VIII.

II. RELATED WORK

There are works such as [5], [6], and [7], which use medical images, such as images captured from colonoscopy videos, endoscopies for the diagnosis of different gastrointestinal diseases. If we refer specifically to the diagnosis of conditions that affect the lungs. We have as a reference [8], where they used images to classify different diseases that occur in the lungs, such as pneumonia, sarcoidosis, and cancer, obtaining in each case more than 78% accuracy. Another similar work is [9], where they classify the two types of pneumonia, either of viral or bacterial origin; for this, the authors divide their methodology into three steps: first, they segment the critical part of the lung, that is, the left and right area of the lung, then they extract the characteristics of each image making use of transfer learning, and finally they use those characteristics for binary classification using the SVM algorithm.

The first works for this task used less amount of data since it was not available due to the virus's recent appearance. Among the most used techniques are convolutional neural networks (CNN). An excellent example of Convolutional Neural Networks' application is the work [10], which could obtain a precision of 85% despite the data limitation. It should be noted that this work was only based on one model, SqueezeNet, compared to the proposed work that is used up to 6 models.

Like [11] and [10], the works are both based on convolutional neural networks, but at this time, due to the little data available, data augmentation was used to have more training data. The work [11] divides work into three phases: a data augmentation phase, the second feature extract phase using already trained CNN models based on transfer learning, and finally, a classification phase.

In other works referred to as [12], carried out a few weeks after starting this pandemic, due to the lack of data, they used algorithms such as Generative Adversarial Networks (GAN) to generate images with positive Covid-19 cases. These data were used to train models, validate them, and diagnose the disease from the generated models. Despite using a GAN instead of data augmentation, the final precision obtained was 77%.

The use of some automatic hyperparameter optimization approach is essential here. The CNN needs adjustments of several hyperparameters that directly affect the model's performance. In-state of the art, they recommend using Bayes optimization, having as its main characteristic the consideration of past iterations, which is why it was chosen to follow this approach. However, there are tools and libraries such as Auto-Weka [13], Auto-Keras [14], and Google Vizier [15] that promise this automatic optimization of hyperparameters. But they do not have the flexibility to consider, for example, *batch_size*, pre-trained models, or some other characteristic that could be considered as hyperparameters.

Taking as reference the work [16], wherein the same way it uses this Bayes approach for the automatic optimization of hyperparameters considering the pre-entered models and the

descent of gradients' optimization function. However, recently in work [17], they consider that the Batch Size is an essential hyperparameter in the classification of medical images using convolutional networks. They conclude that a large batch size does not necessarily produce better accuracy. Inspired by these studies, it was considered as optimization space: pre-trained models, the optimization function of the descent of gradients, learning rate, momentum, and batch size as hyperparameters.

III. METHODS AND MATERIALS

This section describes how the data were collected and the sources from which they were obtained, and their characteristics. Besides, this work's theoretical concepts are explained in a didactic way, such as the deep learning algorithms, the transfer learning process, and the automatic hyperparameter optimization algorithm.

A. Dataset

For the present study, we collected and used a dataset of chest X-ray radiography (CXR) images acquired from various publicly available medical repositories [18] [19] [20] hosted on Kaggle. At the beginning of the pandemic, these data were scarce, wherein most of the studies used data augmentation to achieve a greater amount of data, both for training and validation.

This repository was developed by a team of researchers from the University of Qatar, Doha, Qatar, Dhaka, Bangladesh, and their collaborators from Pakistan and Malaysia in collaboration with doctors. They created a chest X-ray images database for COVID-19 positive cases along with Normal and Viral Pneumonia Images. COVID data is collected from different public access data sets, online sources, and published articles. All images are in Portable Network Graphics (PNG) file format and 1024 * 1024 pixels and 256 * 256 pixels. As shown in Fig. 1, contain images for the two class (Covid-19 and Non-Covid-19).

The dataset used in this study includes a total of 5,641 2D X-ray images in the posteroanterior (PA) view of the chest. There was an unbalanced distribution of classes, 1300 images labeled as covid, and 4341 as non-covid. In order not to bias our model, 1,300 images were randomly extracted from each class, specifically in the non-covid surplus class as shown in Figure x, contain images for the two class (Covid-19 and Non-Covid-19). The test dataset used in this work was collected by [21], which is intended to simulate the real world. It is also necessary to highlight that this data is not within the training data. This test dataset contains 5000 non-covid images and 184 images categorized as covid-19.

B. Convolutional Neural Networks (CNN)

The convolutional neural network (CNN) is a deep learning neural network class. In short, think of CNN as a machine learning algorithm that can take an input image, assign importance (weights and learnable biases) to various aspects/objects in the image, and differentiate one from the other. CNN works by extracting features from images. Any CNN consists of the following:

- The input layer is a grayscale image.

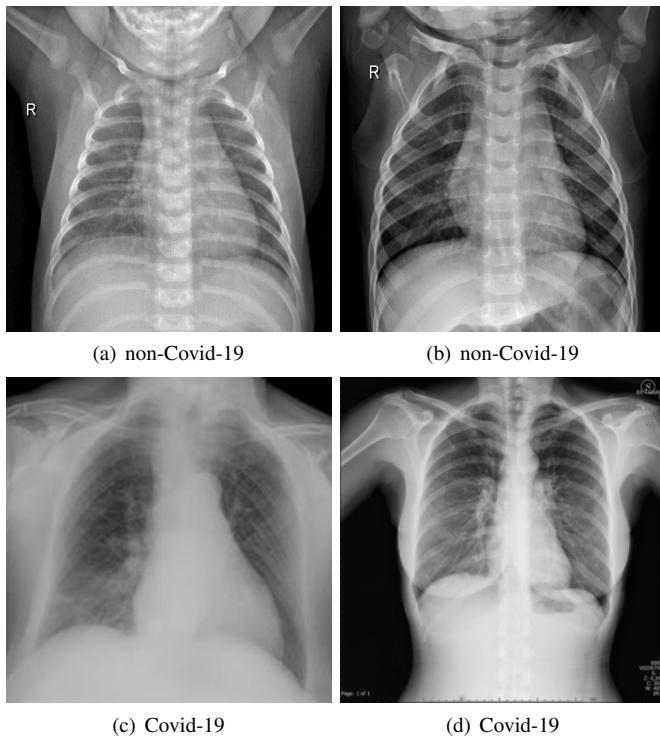


Fig. 1. Example Images from Each Class Covid-19 and non-Covid-19.

- The output layer, which is a binary or multi-class label.
- Hidden layers consisting of convolution layers, ReLU (rectified linear unit) layers, grouping layers, and a fully connected neural network.

It is essential to understand that artificial neural networks, made up of multiple neurons, cannot extract features from the image. This is where a combination of convolution and grouping layers comes into play. Similarly, convolution and grouping layers cannot perform classification, so we need a fully connected neural network.

1) *Convolution*: In the Convolution, product and sums operations are performed between the starting layer and the n filters (or kernel) that generate a characteristic map. The extracted characteristics correspond to each possible location of the filter in the original image. The advantage is that the same filter (= neuron) is used to extract the same characteristic in any part of the input, with this that manages to reduce the number of connections and the number of parameters to train compared to a multilayer network of real connection.

After applying the Convolution, an activation function is applied to the feature maps. The recommended activation function is sigmoid ReLU, selecting a suitable learning rate and monitoring the fraction of dead neurons; it could also be tried with Leaky ReLU or Maxout, but never use logistic sigmoid.

2) *Reduction*: In Reduction, the number of parameters is reduced by keeping the most common characteristics. The way to reduce parameters is done by extracting statistics such as the average or maximum of a fixed region of the characteristics

map; when reducing characteristics, the method loses precision although its compatibility improves.

3) *Sorter*: At the end of the convolutional and Reduction layers, it is often used fully connected layers in each pixel is considered as an individual neuron as in a multilayer perceptron. The last layer of this network is a classifier layer that will have as many neurons as the number of classes to predict.

C. Transfer Learning

Transfer Learning is the act of transferring knowledge from one network to another or, in general, from one model to another as shown in Fig. 2. Convolutional neural networks, as we know, stand out for learning, on their own, to interpret the images that we pass to them. In other words, they are experts in creating high-quality features. For this reason, pre-trained convolutional neural networks are, in many cases, the ideal feature extractors. We give the definitions of “domain” and “task”, respectively.

A domain D consists of two components: a feature space X and a marginal probability distribution $P(X)$, where $X = x_1, \dots, x_n \in X$. For example, if our learning task is to classify documents, each term is considered a binary characteristic. Hence, X is the space of all vectors of terms, x_i is the i th vector of terms corresponding to some documents, and X is a particular learning sample. In general, if two domains are different, then they may have different feature spaces or different marginal probability distributions.

Here is a unified definition of transfer learning. Definition 1 (Transfer Learning). Given a source domain D_s and a learning task T_s , a target domain D_T , and a learning task T_T , transfer learning aims to help improve the learning of the target predictive function f_T in D_t using the knowledge in D_s y T_s , where $D_S \neq D_T$, or $T_s \neq T_T$.

In the above definition, a domain is a pair $D = X, P(X)$. Therefore, the $D_S \neq D_T$ condition implies that $X_S \neq X_T$ or $P_S(X) \neq P_T(X)$. For example, in our document classification example, between a set of source documents and a set of destination documents, the term’s characteristics are different between the two sets (they use different languages), or their marginal distributions are different.

Similarly, a task is defined as a pair $T = \cdot, P(Y|X)$. Therefore, the $T_S \neq T_T$ condition implies that $Y_S \neq Y_T$ or $P(Y_S|X_S) \neq P(Y_T|X_T)$. When the destination and source domains are the same, $D_S = D_T$, and their learning tasks are the same, that is, $T_S = T_T$, the learning problem becomes a traditional machine learning problem. When the domains are different, then 1) the feature spaces between the domains are different, that is, $X_S \neq X_T$, or 2) the feature spaces between the domains are the same, but the marginal probability distributions between the domain data are different; that is, $P(X_S) \neq P(X_T)$, where $X_{S_i} \in X_S$ and $X_{T_i} \in X_T$. For example, in our document classification example, case one corresponds to when the two sets of documents are described in different languages. Case two may correspond to when the source domain documents and destination domain documents focus on different topics. Given the specific domains D_S and D_T , when the learning tasks T_S and T_T are different, then

1) the label spaces between the domains are different, that is, $Y_S \neq Y_T$, or 2) the conditional probability distributions between the domains are different; that is, $P(Y_S|X_S) \neq P(Y_T|X_T)$, where $Y_{S_i} \in Y_S$ and $Y_{T_i} \in Y_T$. In our document classification example, case 1 corresponds to the source domain has binary document classes, while the destination domain has ten classes to classify the documents. Case 2 corresponds to the situation where the source and destination documents are very unbalanced in terms of the user-defined classes.

Furthermore, when there is some relationship, explicit or implicit, between the two domains' feature spaces, we say that the source and destination domains are related.

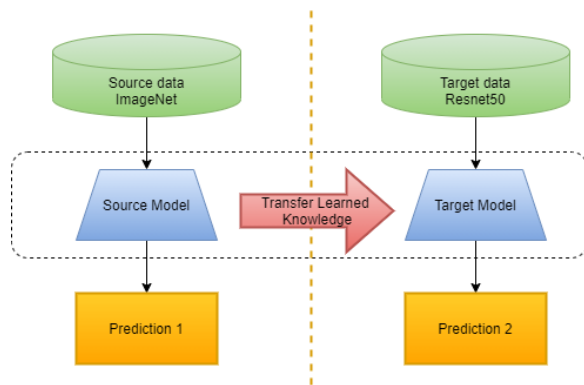


Fig. 2. Transfer Learning Workflow.

D. Bayes Optimization Algorithm

Bayesian optimization is a technique used to optimize an objective function $f(x)$, also called latent or underlying [22]. Its application is made in scenarios where the observations have a high cost, can have noise, and there is no expression for $f(x)$. Given these three characteristics, the objective is to obtain the values that, in addition to providing the most significant amount of information, minimize the objective function in the least possible number of observations.

Scenarios can be proposed in artificial intelligence, such as synthesizing a molecule with certain characteristics, where each evaluation may require a real costly experiment in time and money. In this case, we want to obtain the best parameter configuration, given by a vector x , that obtains the smallest error in $f(x)$. Therefore the mathematical expression that collects this would be the following

$$x_* = \min f(x) \quad (1)$$

Where x is the input value, which minimizes the underlying function $f(x)$, and X is the feature space where it is being optimized. If X is chosen wrong, $f(x)$ is not optimized correctly. In order to effect this minimization, a trade-off will be made between exploiting promising solutions and exploring unknown areas of the entry space. To achieve this trade-off, the acquisition function will use the mean $\mu(x_n + 1)$ and the covariance $\sigma^2(x_n + 1)$. To thus calculate the expected utility of observing a certain point $x_n + 1$. Therefore, Bayesian optimization is a technique that makes its predictions based

on the belief that one has about the model. This approach to problems achieves better results than a simple random search or grid search since neither of these strategies uses the model to guide the minimization process.

The Bayesian optimization pseudocode follows:

- 1) Set $t \leftarrow 0$ randomly generate the initial population $P(0)$
- 2) Select a promising string set $S(t)$ from $P(t)$
- 3) Build network B using a chosen metric and constraints
- 4) Generate a set of new chains $O(t)$ according to the joint distribution encoded by B
- 5) Create a new population $P(t + 1)$ by replacing some strings of $P(t)$ with $O(t)$ set $t \leftarrow t + 1$
- 6) If the termination criteria are not met, go to (2)

In the end, the idea behind Bayesian optimization is to use the model through the Gaussian process in order to calculate the next best point to evaluate, which is calculated using the acquisition function. In this way, the problem becomes optimizing the acquisition function in each evaluation, the cost of which is considered negligible compared to evaluating in the objective function since it lacks noise and is explicitly counted, making it easier to optimize.

a) *Acquisition function:* The acquisition function is of great importance within Bayesian optimization since it regulates the trade-off between exploitation and exploration. There are multiple methods used, such as the probability of improvement (PI), expected improvement (EI), lower confidence bound (LCB), entropy search (ES), and a portfolio of several strategies, the latter usually giving better results. The values returned by the acquisition function usually correspond to the values expected to have a more significant improvement in the optimization task. The first of the listed strategies, probability of improvement, as the name suggests, measures the probability that the following observation is better than the best value obtained so far. Since the values of $f(x)$ are those of a Gaussian posterior distribution, the mathematical expression that defines the probability of improvement is given by

$$\begin{aligned} PI(x) &= P(f(x) \leq f(x^+)) - \varepsilon \\ &= \Phi\left(\frac{f(x^+) + \varepsilon - \mu(x)}{\sigma(x)}\right) \end{aligned} \quad (2)$$

where $\phi(\cdot)$ is the cumulative probability function of standard Gaussian distribution, ε is a regularization constant, and $f(x^+)$ is the best value obtained up to that moment. The problem with this strategy is that it is pure exploitation; once there is a point with a good result, it directs the search only in the vicinity of that point, so it tends to stay local minimums. To also consider unexplored areas, the focus has to be changed to maximize the expected improvement. In order to achieve this, you must first define what the improvement is so that following, you obtain:

$$I(x) = \max\{0, f(x^+) - f(x_n + 1)\} \quad (3)$$

In which $I(x)$ will only have positive values when the evaluation at the new point $x_n + 1$ is less than the previous lower value $f(x^+)$. In order to continue with the calculation of the expected improvement, it must be taken into account that a Gaussian defined the values of $f(x)$, so the probability of improving $I(\cdot)$ in the posterior distribution defined by $\mu(x)$ and $\sigma^2(x)$, is obtained by solving the following integral.

$$\Xi(I) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{f(x^+) + \xi - \mu(x) - I^2}{2\sigma^2(x)}\right) dI \quad (4)$$

where $I = I(x)$, gives as a result, the expected improvement.

$$\begin{cases} EI(x) = (f(x^+) + \xi - \mu(x)\Phi(Z) + \sigma(x)\phi(Z)) & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \quad (5)$$

where

$$Z = \frac{f(x^+) + \xi - \mu(x)}{\sigma(x)} \quad (6)$$

Where $\mu(\cdot)$ is the probability density function of a standard Gaussian distribution.

In both equations shown above, the regulation constant ξ , which is added to $f(x^+)$, is used to specify that there could be a value higher than the best found.

Finally, we will discuss the acquisition function that uses the lowest confidence limit (or the highest, if it is being maximized). This function is optimistic about the variance $\sigma(x)$. Next, the mathematical expression of this strategy is presented

$$LCB(x) = \mu(x) - \nu\sigma(x) \quad (7)$$

where ν is a constant called kappa. Note that $\nu \geq 0$ must be met (since we want to minimize).

IV. METHODOLOGY

As seen in Fig. 3, the present work shows the architecture of the proposed MKCovid-19 workflow. It is made up of two main components of Training and Inference. Each element consists of several sub-components. The training component has a data separation structure, followed by data pre-processing and the sub-component of automatic hyperparameter optimization. Together, they make up the flow training and search for the best model considering the different configurations.

A. Training Component

The flow and procedures to train our MKCovid-19 model for the diagnosis of Covid-19 are detailed, which will depend on the order in which they are performed.

1) *Split Dataset*: MKCovid-19 uses 75% data for training and 25% for validation. The *validation data* is used to avoid overfitting the models. Besides, *test data* is also used to evaluate the model's performance in real situations, that is, the generalizability of the model.

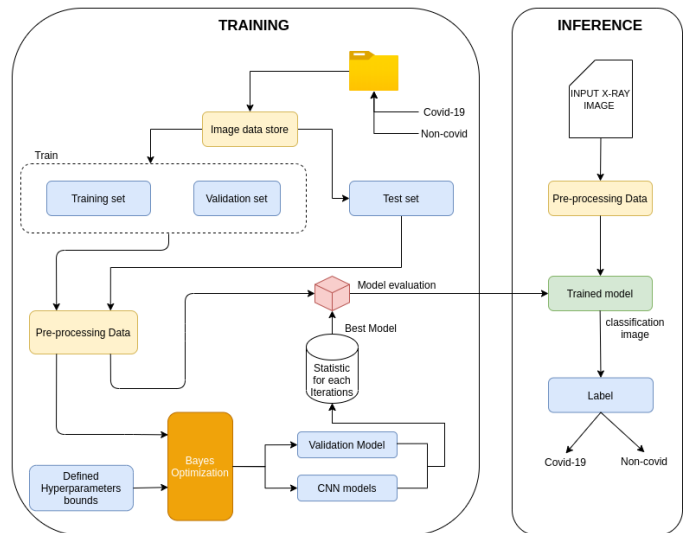


Fig. 3. Architecture of the Proposed MKCovid-19 Model.

2) *Data Pre-Processing*: The input data are prepared following the pre-processing pattern used in training the pre-trained models used in this work. Specifically, the images' size and normalization are converted to 224 * 224 pixels, being the standard size used by the popular models of convolutional neural networks (CNN).

3) *Automatic Optimization of Hyperparameters*: This sub-component receives a search space that contains the values that will not be optimized, such as iterations number and epoch number. It also receives all the hyperparameters that will be optimized, consisting of the learning rate, the batch size, momentum, pre-entered models, and gradient optimizers. At the end of each iteration executed, the statistical data is saved in a CSV file. These data are the values of each hyperparameter chosen by the Bayes optimizer, the loss of each model, and the iteration number. Subsequently, the best configuration is chosen based on the lowest loss obtained. With these configurations of the recovered hyperparameters, the final model is trained to evaluate it with the test data not seen by the training model.

B. Inference Component

The final model generated in the training component is used in this component. For any image consulted, it goes through the same data pre-processing. Finally, the model manages to classify the input image as Covid-19 or Non-Covid-19. This inference component can now be consumed and used by anyone, such as a medical specialist.

V. EXPERIMENTAL SETUPS

As mentioned in the previous sections, one of the main objectives is to provide information on which pre-trained model and hyperparameters can make it possible to obtain more efficient models in the classification of clinical images, specifically in detecting people with Covid-19. This section details the procedures, definitions, and metrics used to carry out our experiments to achieve this objective.

A. Separation of Data into Different Sets

To evaluate the precision of the resulting models trained using transfer learning for clinical data, specifically images of lungs compromised by Covid-19. We want to identify how much it affects the amount of training data, starting from small datasets, up to considerable amounts, in this case, 2,600 images. This experiment is evaluated with the validation dataset and the test dataset.

For this experiment, from the total of 2,600 selected datasets, datasets of different sizes were randomly drawn, such as 150, 300, 500, 1000, 1500, 2000, and 2600 (total data). The complete optimization procedure was executed for each data set, detailed in the Subsection 3.

B. Definition of the Optimization Space

A general non-optimized configuration was defined, with 50 epochs for each iteration and a total of 50 optimization iterations. The hyperparameters search space contains learning rate, momentum, pre-trained models, the gradient descent optimization function, and the batch size. We can observe these hyperparameters in Table I, where it is also shown for continuous data, the minimum, maximum value, and the increment. For discrete data, the set of values is shown.

TABLE I. DEFINE SEARCH SPACE FOR BAYES OPTIMIZATION

Hyperparameters	Min Value	Max Value	Step Value
Batch Size	5	100	5
Momentum	0	1	0.1
Learning Rate	log(0.01)	log(0.02)	-
Pre-Trained Models	Resnet18, Resnet50, Googlenet, Vgg16, Squeezenet, Densenet		-
Optimizer	SGD, Adam, RMSprop, Adagrad, Adadelata, Adamax		-

C. Evaluation Metrics

To evaluate the performance of the resulting model, we have used some metrics that recommend state of the art, such as Accuracy (Acc), precision (Pe), specificity (Sp), sensitivity (Se), and F- score. Besides, the count indices are reported, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) that are used to calculate the metrics as mentioned above, which also allow us to have better analysis and understanding of them.

$$1) \text{ Accuracy (Acc): } \frac{t_p+t_n}{t_p+t_n+f_p+f_n}$$

$$2) \text{ Precision (Prc): } \frac{t_p}{t_p+f_p}$$

$$3) \text{ Specificity (Spc): } \frac{t_n}{t_n+f_p}$$

$$4) \text{ Sensitivity (Sen): } \frac{t_p}{t_p+f_n}$$

$$5) \text{ F1-Score (F1): } \frac{2t_p}{2t_p+f_p+f_n}$$

D. Implementation Details

To implement and evaluate our proposed model, Google Colab [23] was used, which provides an Intel (R) Xeon (R) CPU @ 2.30 GHz, a 12 GB system memory, and a 16 GB Tesla P100 GPU. The implementation of this model was done in the Python 3.5² language. We used the pre-trained models available from the Pytorch library [24]. For the Bayesian optimization of hyperparameters, the Hyperopt library[25] was used explicitly to optimize minimization. All the data used and the codes developed for this work are available for future study in the repository³.

VI. EXPERIMENTAL RESULTS

In this section, an analysis is made of the hyperparameters' values obtained throughout the Bayesian optimization iterations. The frequency of use of each hyperparameter is analyzed together with their selection over time. Finally, the results are presented based on the metrics mentioned in Subsection V-C, comparing the models obtained with the validation data and test data.

A. Analysis of the Use of Pre-trained Models and Gradient Optimizers

In Fig. 4, the number of times that the optimizer used a pre-trained model is observed. The frequent use could indicate that a better model for classification has been obtained using that pre-entered model. The result of executing the optimization for each of the datasets shows us that each one chose different pre-trained models, specifically for the data sets considered small, such as 150, 300, 500, and 1000. This behavior can be explained that when trying to minimize the loss in each iteration by choosing a pre-trained model, overfitting occurred due to the small training data.

On the contrary, if we analyze the frequency of the data sets with more data in this work, that is, 1500, 2000, and 2600 (totality), we can observe that the Densenet model is used more frequently. However, the set of 2600 data also has a similar high usage frequency to the Squeezenet model. These results could indicate that our Bayes optimizer could have decided and/or modeled that Densenet is the best pre-trained model for this problem. Furthermore, we conclude that the Vgg16 model is not recommended for this type of application.

Similarly, in Fig. 5, it is shown that the best optimizer for gradient descent for almost all data sets was Adamax. The difference in frequency of use compared to the other optimizers is considerable. This statistic seems to verify the theory and recommendation on using this optimizer by state of the art.

B. Analysis of the Behavior of the Hyperparameter Distributions

Fig. 6 show the sampling distributions and the search choice of the best hyperparameters for learning rate, batch size, and momentum for each dataset. It is observed that for the Learning Rate, the curves have a similar shape and similar values, independent of each data set's data size. These curves

²<https://www.python.org/>

³<https://github.com/kvvaldez/MKCovid-19>

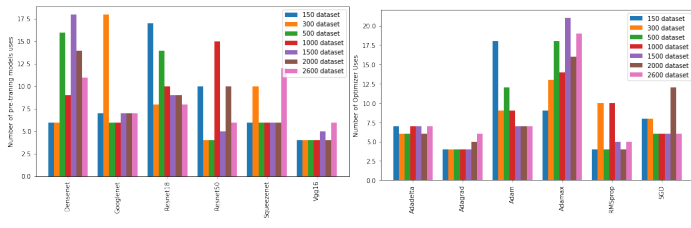


Fig. 4. figure
Frequency of use of Pre-trained Models During Bayesian Optimization.

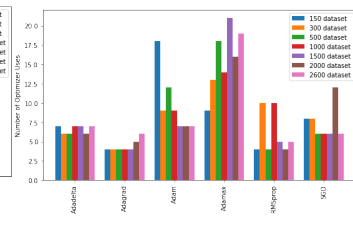


Fig. 5. figure
Frequency of use of Gradient Optimizers During Bayesian Optimization.

have a similar shape to the probability sampling curve but with a range variation of the x-axis. This hyperparameter's best values in the present work range from $< 0.01, 0.035 >$. This range of values could help us define the initial values for a better fit and explore the smallest search space.

For the batch size curve, it is shown that the curves of different data sizes differ from probabilistic sampling, but these curves coincide that it is advisable to use small batch size data in the range of $< 8, 18 >$.

Also, it should be remembered for the momentum curve that this hyperparameter is only used by the SGD and RMSprop gradient optimizers. The momentum figure shows two ridges in the curves in the experiments: $< 0, 0.03 >$ and the other $< 0.8, 0.99 >$, which is close to one. It is necessary to consider that hyperparameter ranges found during Bayesian Optimization are not necessarily better for the test set, only that they produce less loss in the validation data.

Finally, Fig. 7 and Fig. 8 show how the use of pre-trained models and gradient optimizers change over time throughout Bayes optimization. The figure is shown all iterations for each dataset, where it can be observed before iteration 20, many elements belonging to our search space are tested. Later, starting at iteration 20, the optimizer takes a limited number of specific gradient optimizers and pre-trained models. Similarly, in Fig. 10, for each dataset, accuracy was maximized. Although it is observed that the optimization has been minimal, this happens because good results are obtained naturally, at least for this problem. This difference of 1 or two points in the accuracy can mean much gain when using these models to use covid-19 diagnosis and consequently save lives.

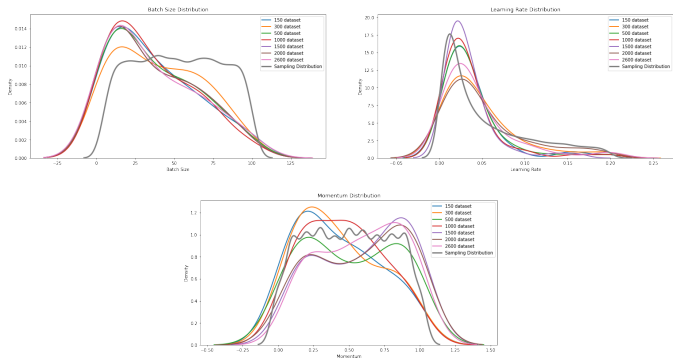


Fig. 6. Sampling Distribution and the Distribution of Learning Rate, Momentum, and Batch Size for the Different Dataset Sizes.

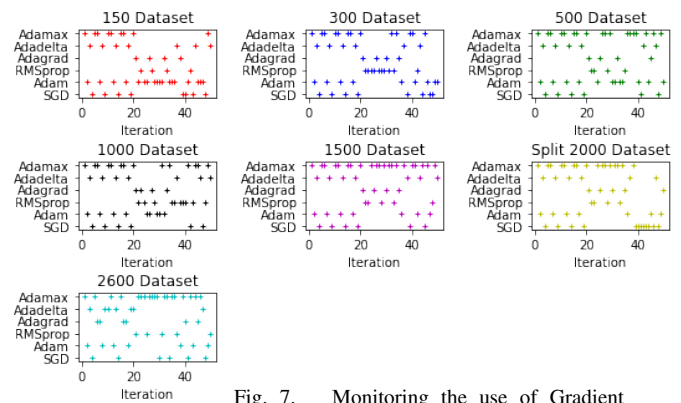


Fig. 7. Monitoring the use of Gradient Optimizer Throughout the Optimization.

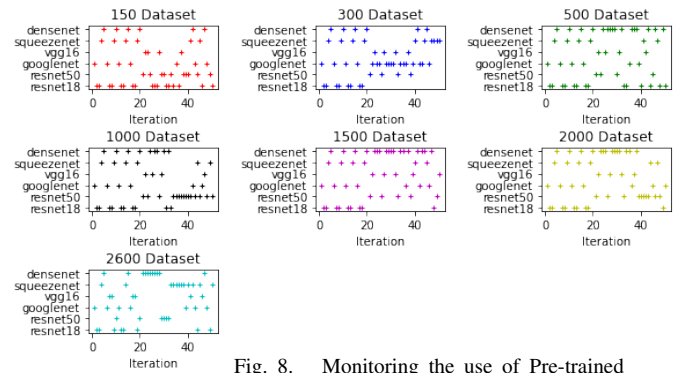


Fig. 8. Monitoring the use of Pre-trained Models Throughout the Optimization.

C. Classification Results and Discussion

Fig. 9 shows the best models' validation precision for each data set. To obtain the best model, all the hyperparameter values were ordered in increasing order according to the loss obtained in each iteration. Then the first best hyperparameter configuration was extracted for each data set.

It is also observed in Fig. 9 the obtaining of models above 94% precision. It seems to confirm that using transfer learning works well for any data amount. Also, from the data set 1000 onwards, a very similar precision was obtained, which is above 97%. The models' performance validation throughout the optimization was carried out with data that simulates a real-world situation. That is, more people think they have Covid-19 without having it, compared to people who have it. This is because their symptoms are not only symptoms of a single disease. For this simulation, the test data was not seen in the training. This data consists of 183 images tagged with Covid-19 and 500 images of people without Covid-19.

The result obtained with the test data is shown in Table II, using a threshold of 0.4. It is observed that low accuracy was obtained for '150 dataset' because there are many false negatives. On the contrary, for '300 dataset', an accuracy of 88% was obtained, which could be considered a good value, but compared to 98% in the validation, it is a considerable difference. This difference in accuracy values could verify the hypothesis mentioned within the Subsection VI-A where the model cannot learn or generalize its learning because little data was used for training. From the "500 dataset" data set, accuracy values similar to the validation one are obtained,

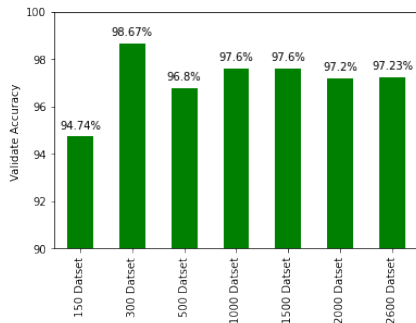


Fig. 9. Validate Accuracy of the Best Models from Each Data Set.

obtaining the best accuracy for data totality (2600 dataset). This high accuracy value is due to the fact that the model was trained with a more significant amount of data, which manages to learn and generalize all the characteristics of an image with a diagnosis of Covid-19.

TABLE II. RESULTS FOR TEST DATASET, BASED ON DIFFERENT METRICS.

DATASET	FN	FP	TN	TP	F1	ACC	PRC	SEN	SPC
150 dataset	223	3	277	180	0.614	0.67	0.98	0.98	0.55
300 dataset	59	23	441	160	0.80	0.88	0.87	0.87	0.88
500 dataset	41	10	459	173	0.87	0.93	0.95	0.95	0.92
1000 dataset	10	12	490	171	0.94	0.97	0.93	0.93	0.98
1500 dataset	11	16	489	167	0.93	0.96	0.91	0.91	0.98
2000 dataset	6	13	494	170	0.95	0.97	0.93	0.93	0.99
2500 dataset	11	5	489	178	0.96	0.98	0.97	0.97	0.98

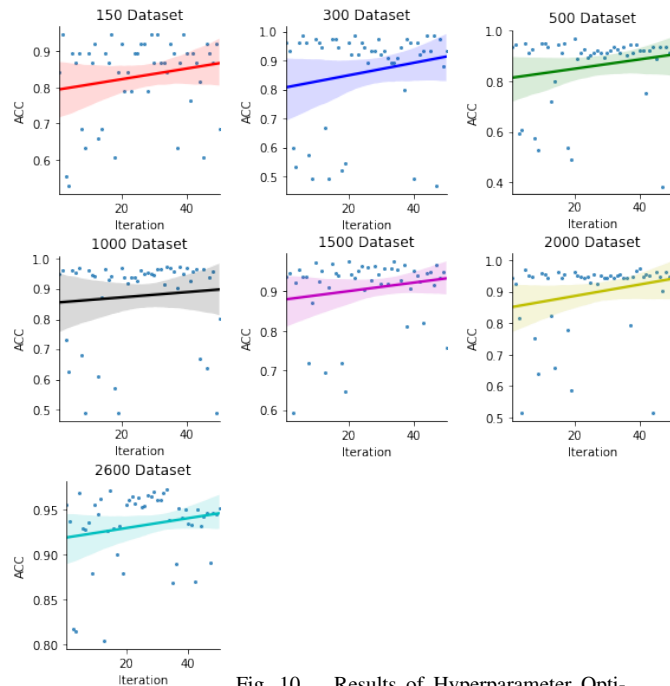


Fig. 10. Results of Hyperparameter Optimization across Iterations.

VII. CONCLUSION

A detailed study was conducted, and a workflow was presented for the automatic selection of hyperparameters for the classification of images of people with Covid-19 using the

Pytorch library. Several experiments were carried out, within which the hyperparameters were optimized for each data set with different amounts of data. The pre-trained models were included as a hyperparameter in order to try to know and recommend which models best fit this type of medical image classification model. For this reason, we conclude the best fit of the pre-entered models towards our problem is highly variable according to the amount of data. If we consider the results obtained with relatively large data sets, the Densenet models and Resnet in general, provide us with accurate and robust models. As a result of this study, we made available the best hyperparameters of the best model obtained with an accuracy of 98%, which will help to quickly detect people with Covid-19 with high sensitivity, using chest X-ray images. Besides, due to the analysis seen in the VI-B sections, we recommend smaller search spaces for this problem, which will surely be very useful as a starting point for people who want to train a model similar to ours. We then conclude that Bayesian optimization is an effective strategy to increase transfer learning use cases.

VIII. FUTURE WORKS

As a first future work, we recommend doing more iteration in the optimization component to analyze better and conclude the behavior of the hyperparameters considered in the search space in this work. We recommend testing the other pre-trained models not addressed in this study regardless of the image size they were trained in since only models with image size 224 * 224 pixels were used. Finally, it is recommended to validate this optimization workflow to classify other diseases that are not necessarily Covid-19 images.

REFERENCES

- [1] Majid Nour, Zafer Cömert, and Kemal Polat. A novel medical diagnosis model for covid-19 infection detection based on deep features and bayesian optimization. *Applied Soft Computing*, 97:106580, 2020.
- [2] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020.
- [3] Mohammad Rahimzadeh and Abolfazl 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 in Medicine Unlocked*, 19:100360, 2020.
- [4] Antonios Makris, Ioannis Kontopoulos, and Konstantinos Tserpes. Covid-19 detection from chest x-ray images using deep learning and convolutional neural networks. In *11th Hellenic Conference on Artificial Intelligence*, pages 60–66, 2020.
- [5] Konstantin Pogorelov, Michael Riegler, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Carsten Griwodz, Peter Thelin Schmidt, and Pål Halvorsen. Efficient disease detection in gastrointestinal videos—global features versus neural networks. *Multimedia Tools and Applications*, 76(21):22493–22525, 2017.
- [6] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017.
- [7] Michael Riegler, Mathias Lux, Carsten Griwodz, Concetto Spampinato, Thomas de Lange, Sigrun L Eskeland, Konstantin Pogorelov, Wallapak Tavanapong, Peter T Schmidt, Cathal Gurrin, et al. Multimedia and medicine: Teammates for better disease detection and survival. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 968–977, 2016.

- [8] Qiao Ke, Jianshe Zhang, Wei Wei, Dawid Potap, Marcin Woźniak, Leon Koźmider, and Robertas Damasevicius. A neuro-heuristic approach for recognition of lung diseases from x-ray images. *Expert Systems with Applications*, 126:218–232, 2019.
- [9] Xianghong Gu, Liyan Pan, Huiying Liang, and Ran Yang. Classification of bacterial and viral childhood pneumonia using deep learning in chest radiography. In *Proceedings of the 3rd International Conference on Multimedia and Image Processing*, pages 88–93, 2018.
- [10] Matteo Polsinelli, Luigi Cinque, and Giuseppe Placidi. A light cnn for detecting covid-19 from ct scans of the chest. *Pattern Recognition Letters*, 140:95–100, 2020.
- [11] Sakshi Ahuja, Bijaya Ketan Panigrahi, Nilanjan Dey, Venkatesan Rajinikanth, and Tapan Kumar Gandhi. Deep transfer learning-based automated detection of covid-19 from lung ct scan slices. *Applied Intelligence*, 51(1):571–585, 2021.
- [12] Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Randgan: Randomized generative adversarial network for detection of covid-19 in chest x-ray. *arXiv preprint arXiv:2010.06418*, 2020.
- [13] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855, 2013.
- [14] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1946–1956, 2019.
- [15] Daniel Golovin, Benjamin Solnik, Subhdeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017.
- [16] Rune Johan Borgli, Håkon Kvale Stensland, Michael Alexander Riegler, and Pål Halvorsen. Automatic hyperparameter optimization for transfer learning on medical image datasets using bayesian optimization. In *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*, pages 1–6. IEEE, 2019.
- [17] Ibrahem Kandel and Mauro Castelli. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT express*, 6(4):312–315, 2020.
- [18] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020.
- [19] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M. Zughair, Muhammad Salman Khan, and Muhammad E.H. Chowdhury. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319, 2021.
- [20] Amith Khandakar Tawsifur Rahman, Dr. Muhammad Chowdhury. Covid-19 radiography database, 2021. <https://www.kaggle.com/tawsifurrahman/covid19-radiography-database>.
- [21] Shervin Minaee, Rahele Kafieh, Milan Sonka, Shakib Yazdani, and Ghazaleh Jamalipour Soufi. Deep-covid: Predicting covid-19 from chest x-ray images using deep transfer learning. *arXiv preprint arXiv:2004.09363*, 2020.
- [22] Daniel Fernández Sánchez. Creating a bayesian optimization tool in python. pages 15–18, 2019.
- [23] Ekaba Bisong. Google colabatory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Springer, 2019.
- [24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
- [25] James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.

Impact of Deep Learning on Localizing and Recognizing Handwritten Text in Lecture Videos

Lakshmi Haritha Medida^{1*}
Research Scholar, CSE
JNTUA, Ananthapuramu
India

Kasarapu Ramani²
Soft Computing Research Centre, Department of IT
Sree Vidyanikethan Engg College (Autonomous)
Tirupati, India

Abstract—Now-a-days, the video recording technologies have turned out to be more and more forceful and easier to utilize. Therefore, numerous universities are recording and publishing their lectures online in order to make them reachable for learners or students. These lecture videos encapsulate the handwritten text written either on a paper or blackboard or on a tablet using a stylus. On the other hand, this mechanism of recording the lecture videos consumes huge quantity of multimedia data in a faster manner. Thus, handwritten text recognition on the lecture video portals has turned out to be an incredibly significant and demanding task. Thus, this paper intends to develop a novel handwritten text detection and recognition approach on the video lecture dataset by following four major phases, viz. (a) Text Localization, (b) Segmentation (c) Pre-processing and (d) Recognition. The text localization in the lecture video frames is the initial phase and here the arbitrarily oriented text on video frames is localized using the Modified Region Growing (MRG) algorithm. Then, the localized words are subjected to segmentation via the K-means clustering, in which the words from the detected text regions are segmented out. Subsequently, the segmented words are pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the Deep Convolutional Neural Network (DCNN). The performance of the proposed model is analyzed in terms of the performance measures like accuracy, precision, sensitivity and specificity to exhibit the supremacy of the text detection and recognition in lecture video. Experimental results reveal that at Learning Percentage of 70, the presented work has the highest accuracy of 89.3% for 500 count of frames.

Keywords—Lecture video; text localization; segmentation; word recognition; Deep Convolutional Neural Network (DCNN)

I. INTRODUCTION

In the recent days, the professional lecture videos are abundant and the number is constantly growing in the web. These lecture videos are motivating the students towards tele-teaching and e-learning [1] [2] [3] [4]. It is more crucial for the students to quickly understand the subject by viewing the video rather than reading the text. The advanced analysis techniques help to automatically collect the relevant metadata from these videos and hence the video lectures are becoming an easiest technique of online course learning [5] [6] [7]. In the MOOCs, it is vital to understand the lecture videos for educational research as it has become synonymous with distance learning. The better understanding of the lecture video lies in the vital cues like the figures, images and text [8] [9] [10] [11]. Among these vital cues, the text is available in

almost all lectures as it can be utilized for variety of tasks like the extracting class notes, generation of keywords, search enabling and video indexing.

In the lecture video, the handwritten text can be on a blackboard or a paper and this text can be written using a stylus on a tablet and displayed on a screen or font rendered text appearing in presentation slides (digital text). These lectures are usually documented with typically positioned cameras. In general, the identification of the text from the presentation slides is a bit easier while compared to the handwritten blackboard text, since they are more legible. On the other hand, the handwritten text recognition is a different beast considering the amount of variations and the character overlaps. The Characters can be small/large, stretched out, swooped, stylized, slanted, crunched, linked, etc. Digitizing handwritten text recognition is extremely exciting and is still far from solved - but deep learning is assisting us in improving the accuracy of the handwritten text recognition. The handwritten blackboard text recognition is additionally challenging and is not legible due to lower contrast, bad illumination, smaller size letter etc. Moreover, detection of the text on the blackboard or paper might be difficult and cluttered, if the lecture over-writes or writes over the figures and equations [12] [13] [14]. The Handwritten Text Recognition (HWR) focuses on the handwritten text in documents and it is practically inherent to complexity in case of different writing styles.

Over the decades, extensive research has been carried out in the field of text recognition on the lecture videos and a variety of methods and algorithms were developed. Word spotting is a key challenge and the majority of the up to date works uses DCNN for learning the features. DCNNs learn the features of the word from dissimilar attribute spaces and are invariant to diverse styles and degradations. Thus, with due interest to handwritten text identification on the lecture video with utmost accuracy, this work focuses on formulating a novel technique by specifically looking into the problems of existing works.

The major contribution of the current research work is highlighted below:

- A novel deep learning based handwritten text recognition approach for video lectures is developed.

*Corresponding Author

- Initially, the text in the collected video frames is localized with Modified Region Growing Approach and these texts are segmented with K-means clustering.
- The segmented words are pre-processed and recognized using DCNN.
- The performance of the proposed model will be analyzed in terms of certain performance measures like accuracy, precision, sensitivity and specificity.

The rest of the paper is organized as: Section II discusses about the literature works undergone under this subject. Section III portrays about the proposed handwritten textual recognition in lecture videos. The resultant acquired with the presented work is discussed in Section IV. Finally, a strong conclusion is given to the current research in Section V.

II. LITERATURE REVIEW

In 2014, Yang *et al.* [15] have developed a novel framework for video text detection and recognition. A Fast Localization-Verification Scheme (FLVS) with the Edge Based Multi-Scale Text Detection (EMS-TD) was constructed in the text detection stage. This algorithm consists of three main steps: text gradient direction analysis, seed pixel selection and seed-region growing. The novel video text binarization algorithm was employed for better text recognition. The potential text candidates were detected with high recall rate by the edge based multi-scale text detector. Then, the detected text lines of the candidate in the video were refined by using image entropy-based filter. Subsequently, the false alarms in the lecture video were discarded by the authors with the help of the Stroke Width Transform (SWT) and Support Vector Machine (SVM). In addition, a novel skeleton-based binarization method was constructed to disconnect text from compound backgrounds in the text recognition phase. The proposed text recognition model in lecture video was evaluated in terms of accuracy using the publicly available test data sets.

In 2018, Poornima & B. Saleena [16] has developed a new technique for successful repossession of the lecture videos from the database using the Correlated Naive Bayes (CNB) classifier. Here, the textual features as well as the image texture were extracted from the key frames with the help of the Tesseract Classifier (TC) and Gabor Ordinal Measure (GOM). The extracted feature dataset encloses three major types of features like the keywords, semantic words, and the image texture. On the basis of the similarity of the features, the authors grouped the video with K-means clustering. Finally, the texts were recognized from the lecture video on the basis of the correlation as well as the posterior probability. The proposed model was compared over the existing models in terms of precision and recall.

In 2018, Kota *et al.* [17] have constructed a Deep Learning based method for handwritten text, math expressions and sketches recognition in the online lecture videos. In the proposed model, the input from the whiteboard lecture video was recorded by the video processing pipeline using a still camera. Then, the summary of the handwritten elements on the whiteboard in the lecture was generated as keyframes over

time. It suffers from the occluded content owing towards the motion of the lecturer. They implied the conflict minimization approach after spatio-temporal content associations with the aim of generating the summary of the key frames. In addition, the Coarse-Grained Temporal Refinement (CTR) was employed to the Content Bounding Boxes (CBB) to detect the variations in the detector output in terms of dissimilarity like the occlusions and illumination.

In 2015, Husain *et al.* [18] have projected a distributed system in which the lecture video frames were stored in the Hadoop's Distributed File System (HDFS) repository. Then, with the help of the HDFS, the processing operations and the highly concurrent images were processed. Further, the MapReduce framework was implied for reading text information as well as for counting the frequent appearance of the words. The proposed text recognition and word count algorithms were tested with the cluster size of 1 and 5 in the Hadoop framework. The resultant of the proposed model confirmed its application in the field of video processing and high-speed image processing.

In 2019, Dutta *et al.* [19] have investigated the efficiency of the traditional handwritten text recognition and word spotting methods on the lecture videos. The dataset was collected from LectureVideoDB having 24 different courses across science, management and engineering. Once the frames were stored, they were pre-trained using the TextSpotter. They localized the words in the video lecture using the deep Fully Convolutional Neural Network (FCNN) and to the output of FCNN; the Non-Maximal Suppression (NMS) was employed to detect the arbitrarily oriented text on the blackboards. Once, the location of the word is identified, the word was recognized using the Convolutional Recurrent Neural Network (CRNN) architecture and Convolutional Recurrent Neural Networks Based Spatial Transformer Network (CRNN-STN). Then, as a novelty they spotted the keywords in the video by extracting the features with two parallel streams of network and label information was concatenated using Pyramidal Histogram of Characters (PHOC) features.

In 2019, Miller [20] has designed a lecture summarization service model by leveraging Bidirectional Encoder Representations from Transformers (BERT) model. The initial contribution of this research work was based on the supervision of lecture transcript and summarizations, which help the users to edit, retrieve and delete the stored items. The second contribution of this research work was an inference from the BERT model with K-Means model in order to produce the embeddings for clustering. Further, on the basis of specified configuration, the summaries for users were generated by BERT model. Finally, the proposed BERT model was compared with the TextRank and the resultant exhibited no golden truth summaries, but there was improvement in the handling context words and was applicable to more lecture videos.

In 2015, Miller *et al.* [21] have constructed a new approach for Automated Video Content Retrieving (AVCR) within large lecture video archives. Initially, the audio lecture was separated from the video and the video was converted into image key-frame using Optical Character Recognition (OCR)

algorithm. Then, from the image, the keywords were extracted using the OCR algorithm. Subsequently, for the video content navigation, a visual guidance was provided by the key-frame detection as well as the automatic video segmentation model. Then, the video OCR was employed on the key-frames and Automatic Speech Recognition (ASR) in order to extract the textual metadata available in the lecture videos. Further, on the basis of the multimedia search diversification method, appropriate answers were collected on the basis of the words. The proposed model had provided more relevant information with more effectiveness to the users.

In 2019, Husain and Meena [22] have introduced a novel method for efficient Automatic Segmentation and Indexing of Lecture Videos (AS-ILV). The proposed model helps in faster reorganization of the specific and relevant content in the online lecture video. In the proposed model, the authors projected the automatic indexing of lecture videos from both slide text and audio transcripts with the extracted topic hierarchies in the video. On the basis of the slide text information, the authors have indexed the videos with higher character recognition rates in an accurate manner. As a novelty, the authors have overcome the problem of high Word Error Rate (WER) transcribed in the video due to the unconstrained audio recording with the semi-supervised Latent Dirichlet Allocation (LDA) algorithm. They have tested the proposed model with Coursera, NPTEL and KLETU classroom videos and the resultant of the evaluation exhibited average percentage improvement in F-Score, while compared to the existing one. Table I summarizes the above-mentioned works along with the methodology, features and challenges.

Regardless of massive amount of works on lecture videos and MOOCs, there are incredibly a small number of which distinctively come across this problem. Among them, the SVT and SVM approach in [1] has high robustness and High recall rate. Apart from these advantages, it requires improvement in the text recognition rate with the aid of the context- and dictionary-based post processing and the text detection result need to be improved with the help of the text tracking algorithms. Further, in CNB, tesseract classifier and GOM [2], the computational time is lower and the precision as well as recall are improved. This technique suffers from retrieval of text from large dataset and hence optimization technique needs to be implied. In [3], the conflict minimization approach gives higher text detection rate. This technique need to handle occlusions and temporal refinement for end-to-end detection of content in video frames. Then, HDFS and MapReduce in [4] is a fault tolerant distributed system and it is cost effective. The memory shortage problems are created by large datasets and hence memory optimization needs to be implied. Moreover, CRNN and CRNN-STN in [5] is Applicable for low resolution and complex images. But, this technique does not use Applicable for low resolution and complex images. A higher trade-off is achieved between the speed and inference Performance by BERT model. But here the automatic extractive summarization is not perfect. Further, OCR algorithm [7] is good in providing the relevant information in a better way and can collect the appropriate answer for the words. As a controversy to these advantages, it too suffers from low recognition rate and high cost. The F-Score is enhanced by LDA algorithm; however, the WER is not removed completely. The research works in this area should focus on one or more of these problems.

TABLE I. FEATURES AND CHALLENGES OF EXISTING LECTURE VIDEO RECOGNITION APPROACHES

Author [citation]	Methodology	Features	Challenges
Yang <i>et al.</i> [15]	SWT and SVM	✓ High recall rate. ✓ High robustness	× Requires improvement in text recognition rate × Requires improvement in text detection result
Poornima & B. Saleena [16]	CNB , tesseract classifier and GOM	✓ Better precision, recall ✓ Minimum computation ✓ Time	× Retrieval can be done with the optimization techniques × Cannot retrieve text from large databases
Kota <i>et al.</i> [17]	conflict minimization approach	✓ Better text detection ✓ Extract semantically meaningful information	× Need to handle occlusions × Need to investigate temporal
Husain <i>et al.</i> [18]	HDFS and MapReduce	✓ Great improvement in time of execution ✓ Cost effective	× Large datasets often creates memory shortage problems × Requires memory optimization
Dutta <i>et al.</i> [19]	CRNN and CRNN-STN	✓ Good contrast. ✓ Easier to recognize	× Need to use recognized text for larger video understanding problems.
Miller [20]	BERT model	✓ High tradeoff between speed and inference Performance ✓ Improves the quality of recognition	× Automatic extractive summarization is not perfect × Difficulty in handling context words
Miller <i>et al.</i> [21]	OCR algorithm	✓ Applicable for large lecture video archives ✓ Collects appropriate answer for the words	× High cost × Do not address the way of retrieving the appropriate information
Husain and Meena [22]	LDA algorithm	✓ Enhances the F-Score ✓ Provides indexing information	× Poor retrieval performance × WER is not removed completely

III. HANDWRITTEN TEXTUAL INFORMATION RECOGNITION

The pictorial depiction of the adopted video handwritten text recognition approach is revealed in Fig. 1. In the presented scheme, a new handwritten textual image recognition approach is developed using an intellectual technique. The presented scheme comprises of four most important stages such as, Text Localization, Segmentation, Pre-processing and Recognition. At first, the video frames are acquired and the text within the video frames is localized using the Modified Region Growing Algorithm. Subsequently, the localized words are subjected to segmentation via the K-means clustering, in which the words from the detected text regions will be segmented out. Subsequently, the segmented words are pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the DCNN. The resultant from DCNN exhibits the recognized textual information from the acquired lecture video.

A. Text Localization

The collected lecturer video frame encompasses the textual and the audio contents [23]. The textual contents in the images $I(i, j)$ are converted into binary image and the white regions are extracted from it. Further, these extracted white regions $I_w(i, j)$ are subjected to region growing approach for localizing the texts. The segmentation of $I_w(i, j)$ occurs via seed points that have to be regularized. A seed point is the commencement stage for region growing and its selection is significant for the segmentation solution. The stages of region growing approach are portrayed in the following steps.

Step 1: The input image $I_w(i, j)$ is split into a huge count of blocks P , in which all the blocks encompass one centre pixel and several vicinity pixels.

Step 2: Then, fix the Intensity threshold (R^i) .

Step 3: For the entire block P , carry out the subsequent course of action in anticipation of the count of blocks that reaches the entire count of blocks for an image.

Step 3(a): Find out the histogram G of all pixels in P .

Step 3(b): The most recurring histogram of the P^{th} block, signified by U^h is fine-tuned.

Step 3(c): Select any pixel, as per U^h and distribute a pixel as seed point with intensity Int_u .

Step 3(d): The adjacent pixels are computed with respect to intensity Int_n .

Step 3(e): Find out the intensity variation of u and n (i.e.) $Dif_{Int} = \|Int_u - Int_n\|$.

Step 3(f): If $Dif_{Int} < R^i$ add the consistent pixel to the region, and hence the region would grow, or go to step 3(h).

Step 3(g): Authenticate if the whole pixels are added to the region. If yes, go to step 2 and then carry out step 3(h).

Step 3(h): Re-estimate the region and discover the new seed points and perform the procedure from step 3(a).

Step 4: Finish the whole process.

The textual information acquired from the region-based approach is $I_{text}(i, j)$, which is fed as input to k-means clustering for cropping the data from the texts.

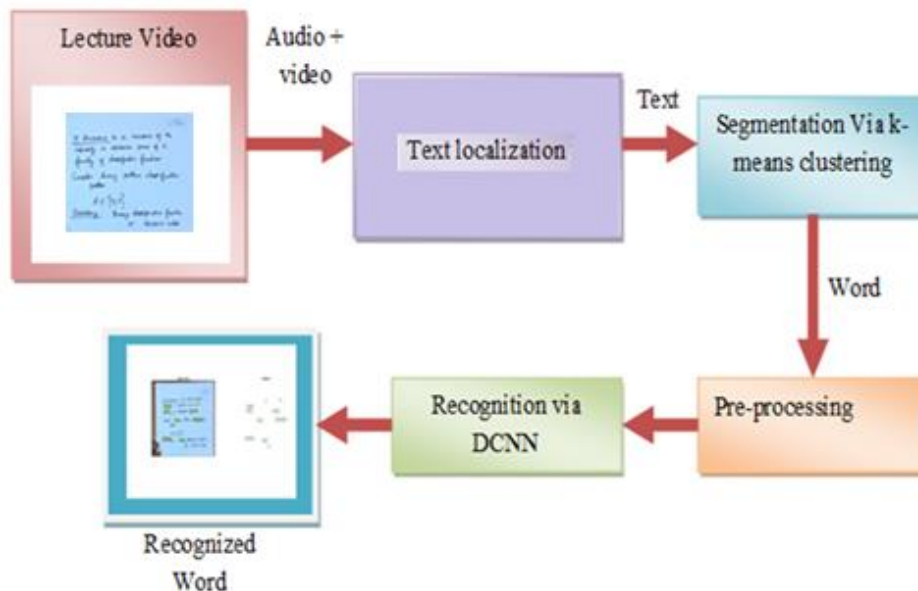


Fig. 1. Proposed Architecture for Handwritten Text Recognition In Lecture Video.

B. Segmentation

Clustering [24] is a scheme that split a group of data into a precise number of sets. A renowned technique among various clustering models is K-means clustering. The words in the texts $I_{text}(i, j)$ are segmented into k count of clusters.

Two separate stages are considered in this algorithm. In the initial stage k centers are selected randomly, in which the k value is previously fixed. Then, in the subsequent stage, every data object is moved to the closest center. Basically, the distance among every object of data and the cluster centers is determined using the Euclidean distance. This process is iterated until the termination criteria happen to a minimum. At the end of segmentation, the words $I_{word}(i, j)$ are extracted and they are subjected to pre-processing.

C. Pre-processing

The gathered words $I_{word}(i, j)$ are pre-processed for enhancing the accuracy of recognition. The steps involved in pre-processing are listed below:

Step 1: Initially, the collected words $I_{word}(i, j)$ are subjected to histogram equalization, which involves transforming the intensity values, i.e. stretching out the intensity range of the image.

Step 2: In order to convert the $I_{word}(i, j)$ image frames into binary image, black-and-white (B&W) images, the adaptive thresholding is used. The binarized image is denoted as $I_{binary}(i, j)$. Further, the weights of these binary images $I_{binary}(i, j)$ are subjected to recognition.

D. Recognition

The weights of the binarized images $I_{binary}(i, j)$ are subjected to recognition via DCNN [25]. Actually, DCNNs are CNNs which encompasses numerous layers and it follows a hierarchical principle. Usually, deep CNNs involve several wholly-connected layers, i.e., layers with dense weight matrix W . To do the recognition process, the outputs are exploited as inputs to a SVM or RF and the output phase can be a softmax function as specified in Eq. (1), in which 1 indicates a column vector of ones.

$$u = \sigma(I_{binary}(i, j)) = \frac{\exp(I_{binary}(i, j))}{1^T \exp(I_{binary}(i, j))} \quad (1)$$

Practically, the entire quantities are turned to be positive by exponential function and accordingly, the normalization assures that the entries of u adds up to 1. Generally, the softmax function is noticed as a multidimensional generalization of sigmoid function deployed in LR. This function termed as softmax is one of the $I_{binary}(i, j)_i$ entries, for instance, if $I_{binary}(i, j)_{b_0}$ is superior over the others, then $I_{binary}(i, j)$ and therefore Eq. (2) is modeled. The above function acts as an indicator amongst the largest entry in x and thus, Eq. (3) is formulated.

$$u_{I_{binary}(i, j)_{b_0}} \approx 1 \text{ and } u_b \approx 0 \text{ for } b \neq b_0 \quad (2)$$

$$\lim_{\alpha \rightarrow \infty} I_{binary}(i, j)^T \sigma(\alpha I_{binary}(i, j)) = \max(I_{binary}(i, j)) \quad (3)$$

In brief, DCNN performs the formulations as specified in Eq. (4)-Eq. (7), in which the output activation function f_x could be softmax, identity, or other function.

$$z^{(0)} = z \quad (4)$$

$$z^{(q)} = \pi(f(W^{(q)}z^{(q-1)})) \quad \text{for } q = 1, \dots, Q_c \quad (5)$$

$$z^{(q)} = f(W^{(q)}z^{(q-1)}) \quad \text{for } q = Q_c + 1, \dots, Q \quad (6)$$

$$I_{binary}(i, j) = f_{I_{binary}(i, j)}(z^{(Q)}) \quad (7)$$

The matrix $W^{(q)}$ consists of $F^{(q-1)} + 1$ columns and $F^{(q)}$ rows with $F^{(0)} = F$ and $F^{(q)}$ for $q > 0$ that is similar to the output count at q^{th} layer. The initial Q_c layers are convolution and the rest ones are wholly-connected. The diagrammatic representation of DCNN is exhibited in Fig. 2.

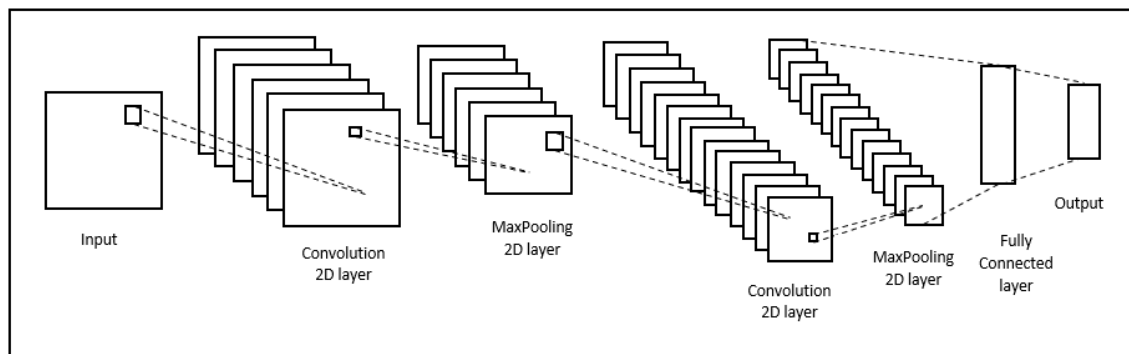


Fig. 2. The Proposed Architecture of DCNN for Handwritten Text Recognition.

IV. RESULTS AND DISCUSSION

A. Simulation Procedure

The proposed video lecture recognition approach is implemented in PYTHON and the resultant acquired is noted. The dataset for the evaluation is downloaded from LectureVideoDB. In addition, two public datasets are utilized for pre-training the word recognition models.

- IAM Handwriting Database: It includes contributions from over 600 writers and comprises of 115,320 words in English.
- MJSynth: This is a synthetically generated dataset for scene text recognition. It contains 8 million training images and their corresponding ground truth words. The sample image collected and its segmented images are depicted in Fig. 3.

This evaluation is accomplished by varying the learning percentage (LP=60, 70) in terms of positive measures. The accuracy, sensitivity, specificity and precision come under the positive measures.

Accuracy(Acc): The accuracy indicates the accurate detection process. The mathematical formula for accuracy is expressed in Eq. (8).

$$Acc = \frac{TrP + TrN}{TrP + TrN + FrP + FrN} \quad (8)$$

PPV or precision: It represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples. It is mathematically shown in Eq. (9).

$$PPV = \frac{TrP}{TrP + FrP} \quad (9)$$

Sensitivity: It is “positive correctly classified samples to the total number of positive samples”. Mathematically, it is expressed in Eq. (10).

$$Sensitivity = \frac{TrP}{TrP + FrN} \quad (10)$$

Specificity: It is the “ratio of the correctly classified negative samples to the total number of negative samples”. This can be mathematically defined in Eq. (11).

$$Specificity = \frac{TrN}{FrP + TrN} \quad (11)$$

where,

TrP - true positive.

TrN - true negative.

FrP - false positive and.

FrN - false negative.

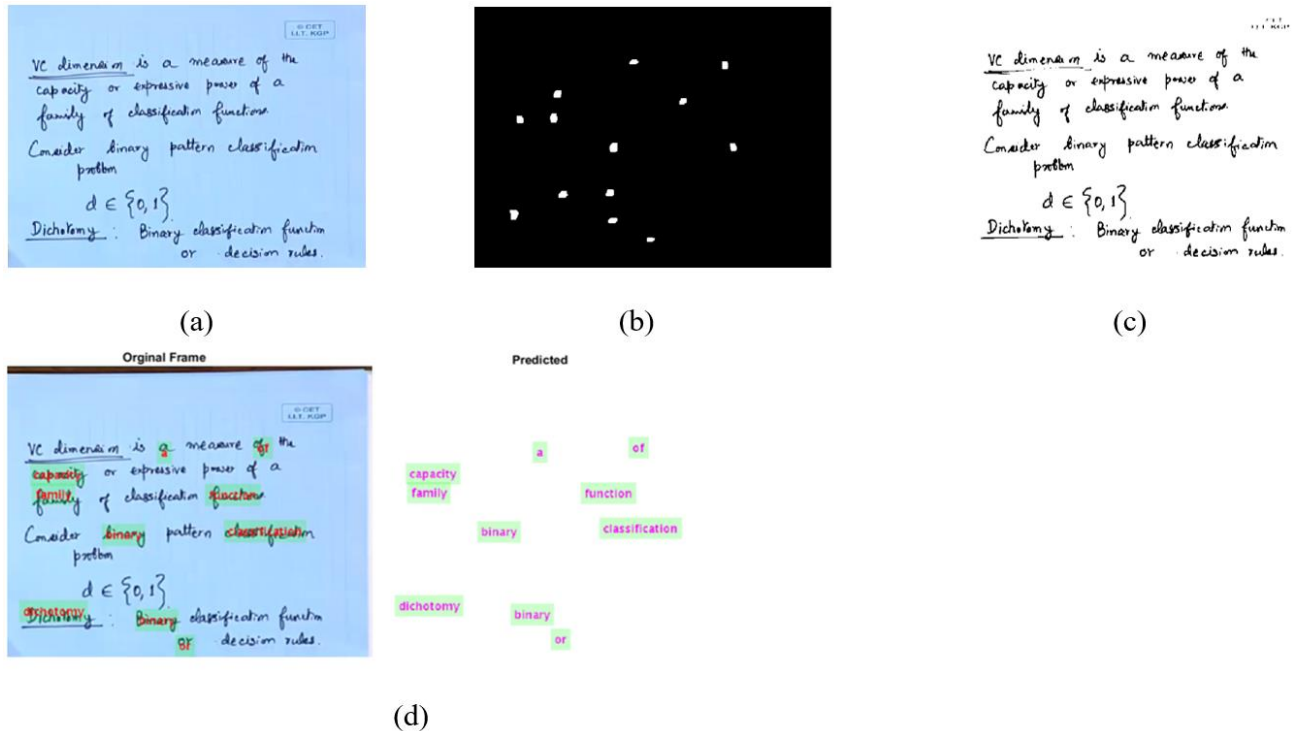


Fig. 3. Sample Images Showing (a) Input Frame (b) Black and White Image Frame (c) Text Localized Frame (d) Recognized Image Frame.

B. Evaluation

The evaluation is done by varying the training percentage (TP). The resultant acquired in terms of positive measures for diverse count frames is shown graphically. Table II and Fig. 4 shows the accuracy of the presented work. It is observed that the presented work has the highest accuracy as 89.3 for 500 count of frames corresponding to both LP =60 and 70. The resultant values of precision acquired are tabulated in Table III and is exhibited graphically in Fig. 5. The highest precision of 95 is obtained for LP=70 for 500 count of frames. The sensitivity of the presented work is highest for both LP=60 and LP=70 at 500 count of frames and the resultants acquired represented in Table IV and Fig. 6. The highest value of sensitivity obtained for the presented work at 500 count of frames is 91.2. The specificity of the presented work for LP=60 and LP=70 is exhibited graphically in Table V and Fig. 7. The specificity of the presented work at LP=70 has the highest value of 91.2 for 500 count of frames and it is higher for LP=70 for every variation in count of frames. The experimental results show that the modern DCNN model shows promising recognition accuracy. On a whole, it is observed the detection rate is higher for LP=70.

TABLE II. EVALUATION ON ACCURACY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Accuracy for LP=60	Accuracy for LP=70
100	84.7	85.6
200	85.6	86.4
300	86.1	87.3
400	87.3	88.1
500	89.3	89.3

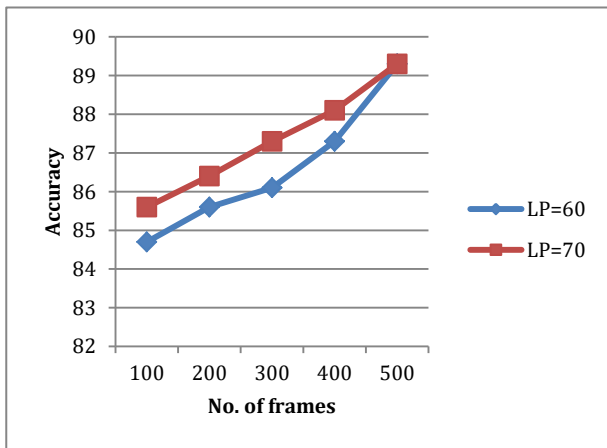


Fig. 4. Evaluation on Accuracy of Presented Work for LP=60 and LP=70.

TABLE III. EVALUATION ON PRECISION OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Precision for LP=60	Precision for LP=70
100	86.5	85
200	87	90
300	93.1	92.5
400	93.9	93
500	94.9	95

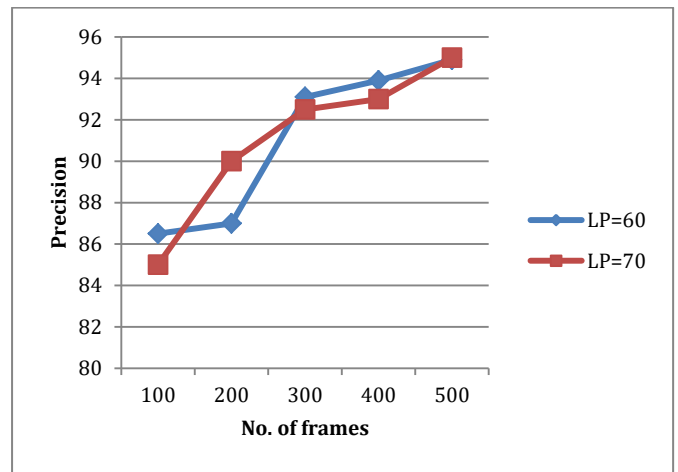


Fig. 5. Evaluation on Precision of Presented Work for LP=60 and LP=70.

TABLE IV. EVALUATION ON SENSITIVITY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Sensitivity for LP=60	Sensitivity for LP=70
100	81.5	86.5
200	87.2	87
300	87.9	88
400	88.5	89
500	91.2	91.2

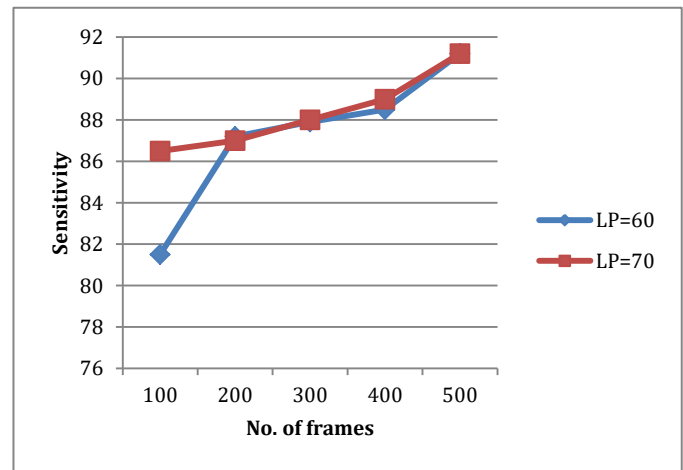


Fig. 6. Evaluation on Sensitivity of Presented Work for LP=60 and LP=70.

TABLE V. EVALUATION ON SPECIFICITY OF PRESENTED WORK FOR LP=60 AND LP=70

No. of frames	Specificity for LP=60	Specificity for LP=70
100	87.1	85.6
200	87.9	88.4
300	88.3	89.1
400	89.9	90.7
500	90.8	91.2

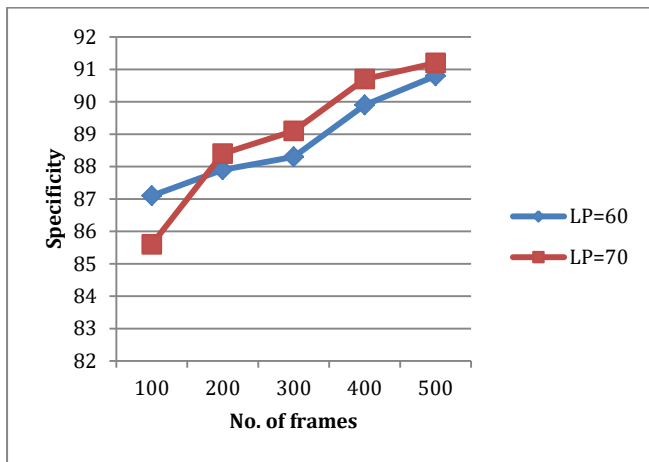


Fig. 7. Evaluation on Specificity of Presented Work for LP=60 and LP=70.

V. CONCLUSION

Although OCR has been considered as a solved problem, Handwritten Text Recognition a crucial component of OCR is still a challenging problem statement. The huge discrepancy in handwriting styles across different people and the poor quality of the handwritten text as compared to the typed or printed text pose substantial hurdles in converting the handwritten text into machine readable text. However, working on this crucial problem is important due to its pertinence in multiple industries such as healthcare, insurance and banking. This paper presented a novel text detection and recognition approach on the video lecture dataset by following four major phases, viz. (a) text localization, (b) segmentation and (c) pre-processing and (d) recognition. In the initial phase, the text localization in the lecture video frames were accomplished using the MRG algorithm. Then, the localized words were subjected to segmentation via the K-means clustering, in which the words from the detected text regions were segmented out. Subsequently, the segmented words will be pre-processed to avoid the blurriness artifacts as well. Finally, the pre-processed words are recognized using the DCNN. The performance of the proposed model is analysed in terms of certain performance measures like accuracy, precision, sensitivity and specificity to exhibit the supremacy of the proposed text detection and recognition in lecture video. Experimental results reveal that at LP=70, the presented work has the highest accuracy as 89.3 for 500 count of frames. In future, some fusion-based DCNN models will be explored for further achieving more accurate detection of handwritten text recognition. Also, a more convincing and robust training could be applied with added preprocessing techniques. We would focus on developing a more comprehensive model with a reduced amount of training time.

ACKNOWLEDGMENT

This work is supported by University Grants Commission (UGC) under Minor Research Project titled "Fast Content Based Search, Navigation and Retrieval system for E-Learning". Project Id: F.No:4-4/2015(MRP/UGC-SERO).

REFERENCES

[1] Siti N. H. Hadie, Anna A. Simok, Shamsi A. Shamsuddin, Jamilah A. Mohammad, "Determining the impact of pre-lecture educational video

on comprehension of a difficult gross anatomy lecture", Journal of Taibah University Medical Sciences, vol.14, no.4, pp.395-401, August 2019. <https://doi.org/10.1016/j.jtumed.2019.06.008>.

[2] Zhongling Pi, Yi Zhang, Fangfang Zhu, Ke Xu, Weiping Hu, "Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures", Computers & Education, vol.128, pp.345-352, January 2019. <https://doi.org/10.1016/j.compedu.2018.10.006>.

[3] Hamidreza Aghababaeian, Ladan Araghi Ahvazi, Ahmad Moosavi, Sadegh Ahmadi Mazhin, Leila Kalani, "Triage live lecture versus triage video podcast in pre-hospital students' education", African Journal of Emergency Medicine, vol. 9, no. 2, pp. 81-86, June 2019. <https://doi.org/10.1016/j.afjem.2018.12.001>.

[4] Alexander R. Toftness, Shana K. Carpenter, Sierra Lauber, Laura Mickes, "The Limited Effects of Prequestions on Learning from Authentic Lecture Videos", Journal of Applied Research in Memory and Cognition, vol. 7, no. 3, pp.370-378, September 2018. <https://doi.org/10.1016/j.jarmac.2018.06.003>.

[5] Aydın Sarıhan, Neşe Colak Oray, Birdal Güllüpinar, Sedat Yanturalı, Berna Musal, "The comparison of the efficiency of traditional lectures to video-supported lectures within the training of the Emergency Medicine residents", Turkish Journal of Emergency Medicine, vol.16, no.3, pp.107-111, September 2016. [10.1016/j.tjem.2016.07.002](https://doi.org/10.1016/j.tjem.2016.07.002).

[6] Marco Furini, "On gamifying the transcription of digital video lectures", Entertainment Computing, vol.14, pp. 23-31, May 2016. <https://doi.org/10.1016/j.entcom.2015.08.002>.

[7] I-Chun Hung, Kinshuk, Nian-Shing Chen, "Embodied interactive video lectures for improving learning comprehension and retention", Computers & Education, vol.117, pp. 116-131, February 2018. <https://doi.org/10.1016/j.compedu.2017.10.005>.

[8] Kep Kee Loh, Benjamin Zhi Hui Tan, Stephen Wee Hun Lim, "Media multitasking predicts video-recorded lecture learning performance through mind wandering tendencies", Computers in Human Behavior, vol. 63, pp. 943-947, October 2016. <https://doi.org/10.1016/j.chb.2016.06.030>.

[9] Andrew T. Stull, Logan Fiorella, Richard E. Mayer, "An eye-tracking analysis of instructor presence in video lectures", Computers in Human Behavior, vol.88, pp. 263-272, November 2018. <https://doi.org/10.1016/j.chb.2018.07.019>.

[10] Juan Daniel Valor Miró, Joan Albert Silvestre-Cerdà, Jorge Civera, Carlos Turró, Alfons Juan, "Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories", Speech Communication, vol. 74, pp. 65-75, November 2015. <https://doi.org/10.1016/j.specom.2015.09.006>.

[11] Rabab El-Sayed Hassan El-Sayed, Samar El-Hoseiny Abd El-Raouf El-Sayed, "Video-based lectures: An emerging paradigm for teaching human anatomy and physiology to student nurses", Alexandria Journal of Medicine, vol. 49, no. 3, pp. 215-222, September 2013. <https://doi.org/10.1016/j.ajme.2012.11.002>.

[12] Feng Wang, Chong-Wah Ngo, Ting-Chuen Pong, "Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis", Pattern Recognition, vol. 41, no. 10, pp. 3257-3269, October 2008. <https://doi.org/10.1016/j.patcog.2008.03.024>.

[13] Karl K. Szpunar, Helen G. Jing, Daniel L. Schacter, "Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education", Journal of Applied Research in Memory and Cognition, vol. 3, no. 3, pp.161-164, September 2014. <https://doi.org/10.1016/j.jarmac.2014.02.001>.

[14] Alendra Lyons, Stephen Reysen, Lindsey Pierce, "Video lecture format, student technological efficacy, and social presence in online courses", Computers in Human Behavior, vol.28, no.1, pp.181-186, January 2012. <https://doi.org/10.1016/j.chb.2011.08.025>.

[15] Haojin Yang, Bernhard Quehl, Harald Sack, "A framework for improved video text detection and recognition", Multimedia Tools and Applications, vol.69, no.1, pp 217-245, March 2014. [10.1007/s11042-012-1250-6](https://doi.org/10.1007/s11042-012-1250-6).

[16] N. Poornima & B. Saleena, "Multi-modal features and correlation incorporated Naive Bayes classifier for a semantic-enriched lecture

- video retrieval system",The Imaging Science Journal, Jan 2018. 10.1080/13682199.2017.1419549.
- [17] B. Urala Kota, K. Davila, A. Stone, S. Setlur and V. Govindaraju, "Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summarization," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, pp. 19-24, 2018. 10.1109/ICFHR-2018.2018.00013.
- [18] M. Husain, Meena S M, A. K. Sabarad, H. Hebballi, S. M. Nagaralli and S. Shetty, "Counting occurrences of textual words in lecture video frames using Apache Hadoop Framework," 2015 IEEE International Advance Computing Conference (IACC), Bangalore, pp. 1144-1147, 2015. 10.1109/IADCC.2015.7154882.
- [19] Kartik Dutta, Minesh Mathew, Praveen Krishnan and C.V. Jawahar,"Localizing and Recognizing Text in Lecture Videos",sep 2019. 10.1109/ICFHR-2018.2018.00049.
- [20] Derek Miller,"Leveraging BERT for Extractive Text Summarization on Lectures",june 2019. arXiv:1906.04165.
- [21] Surabhi Pagar, Gorakshanath Gagare,"Multimedia based information retrieval approach for lecture video indexing based on video segmentation and Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4,no. 12, December 2015. 10.17148/IJARCCE.2015.41273.
- [22] M. Husain and S. M. Meena, "Multimodal Fusion of Speech and Text using Semi-supervised LDA for Indexing Lecture Videos," 2019 National Conference on Communications (NCC), Bangalore, India , pp. 1-6, 2019. 10.1109/NCC.2019.8732253.
- [23] Junhua Chen, Miao Tian, Xingming Qi, Wenxing Wang, Youjun Liu,"A solution to reconstruct cross-cut shredded text documents based on constrained seed K-means algorithm and ant colony algorithm",Expert Systems with Applications, vol.127,pp.35-46,2019. <https://doi.org/10.1016/j.eswa.2019.02.039>.
- [24] Feng Cheng, Shi-Lin Wang, Xi-Zi Wang, Alan Wee-Chung Liew, Gong-Shen Liu,"A global and local context integration DCNN for adult image classification",Pattern Recognition, vol.96,December 2019. <https://doi.org/10.1016/j.patcog.2019.106983>.
- [25] U.-V. Marti and H. Bunke, "The IAM-database: an english sentence database for offline handwriting recognition," IJDAR, 2002.10.1007/s100320200071.

A Vehicle Routing Problem for the Collection of Medical Samples at Home: Case Study of Morocco

Ettazi Haitam¹, Rafalia Najat², Jaafar Abouchabaka³

LaRI Laboratory, Faculty of Sciences
University Ibn Tofail, Kenitra, Morocco

Abstract—This paper aims to solve the problem of sampling and collecting blood and/ or urine tubes from sick people at home via a medical staff (nurse/ caregiver) to the laboratory in an optimal way. To ensure good management, several constraints must be taken into account, namely, staff schedules, patient preferences, the maximum delay time for a blood sample, etc. This problem is considered as a vehicle routing problem with time windows, preference and priority according to urgent cases. We first proposed a mathematical formulation of the problem by using a mixed integer linear programming (MILP) as well as various metaheuristics. Also, we applied this method to a real instance of a laboratory in Morocco (Témara) named Laboratory BioGuich, which gave the most optimal results.

Keywords—Optimization; metaheuristics; vehicle routing problem; allocation and planning; home care

I. INTRODUCTION

For twenty years, the population of Morocco has climbed by ten million according to the 2014 census estimating a considerable increase in the rate of aging people implying an increase in the number of patients with chronic diseases requiring treatment and periodical samples to ensure adequate medical follow-up. This changes the philosophy of doing things that encourage medical structures (hospitals, laboratories, etc.) to diversify their services and thus to move around on their own in order to provide the necessary care and samples for patients.

In general, a pickup and delivery problem (PDP) is a problem where a client can request that a certain quantity of goods and merchandise have to be delivered or collected in their homes or a specific sites. Usually, a couple origin/destination is considered for each product. The author in [1] proposed a bi-objective genetic algorithm in order to resolve the PDP applied on oil distribution. The PDP became the Dial-a-Ride Problem (DARP) when transporting people is concerned instead of merchandise. This variant includes a quality service with the aim of minimizing the wait time and wait return generally used for modeling the transportation issue of disabled and aged people from their homes to the hospital. A more detailed state of the art concerning these problems can be found in the works of [2] and [3].

This article deals more specifically with the problem of collecting and delivering samples during a specific time interval where a qualified caregiver is assigned the task of doing the process.

II. LITERATURE REVIEW

In the context of home medical sampling services, the issue of logistics affects several points, namely the problem of allocation and planning as well as the problem of pickup and delivery.

In the work of [4], [5] and [6], the authors considered the planning and routing of caregivers (nurses) at home and developed a system of spatial decision support. To do this, a heuristic was carried out to build routes for each nurse while limiting itself to the constraint of unavailability of staff and caregivers. Another aspect of the problem was introduced by [7] who proposed a variant of the Dial-a-Ride problem where a team of caregivers is sent simultaneously to the patient in order to perform a required service before the main service (care / blood samples, etc.) such as complete washing of the patient, change of clothing, blood pressure measurement, taking pills... This requires coordination between the various actors (caregivers, nurses...) to provide these services in an optimal way. This same author underlined the importance of the constraints of synchronizations which make it possible to emphasize the priority aspect of the visits.

The author in [8] discussed a vehicle scheduling problem encountered in home care logistics. This involves providing patients with medical devices and home care company drugs, delivering special drugs from hospital to patients, and collecting biological samples and unused drugs and medical devices for patients. The problem can be seen as a particular vehicle routing problem with simultaneous deliveries, pick-up and time windows, with four types of requirements: delivery from the depot to the patient, delivery from a hospital to the patient, pickup from patient location to depot and pickup from patient location to a medical laboratory. In the event that the patient needs medication provided by the hospital, the vehicle should visit the hospital first. Two mixed programming models have been proposed with a genetic algorithm (GA) and a taboo search method (TS). The genetic algorithm is based on a chromosome permutation, a division procedure and a local search. The attributes for assigning routes for patients form the basis of the tabu search method. These approaches are tested on test cases from existing VRPTW benchmarks. References [9] and [10] proposed a vehicle routing problem with time windows, synchronization, precedence and lunch break constraints for home care services in order to add a realistic aspect of the problem by adding the break aspect for caregivers, the authors applied a mathematical formulation solved by a linear solver for small instances, then used various metaheuristics for larger instances.

TABLE I. OBJECTIVES ENCOUNTERED IN THE RECENT PAPERS

	T	C	D	PP	CP	AS
[11]	—					
[12]		—				
[13]		—		—	—	
[14]	—		—			
[6]			—			
[15-16]			—	—		
[17]					—	—
[18]	—			—	—	
[19]	—					
[20]		—			—	—
[8]		—				
[21]		—				
[7]	—					—
[22]		—				
[23]	—					
[24]		—				—
[25]	—					
[26]			—	—		
[27]	—	—	—	—		
[28]	—	—	—			—
[29]	—	—	—			—
[30]	—	—				

TABLE II. COMMON OPTIMIZATION CRITERIA

Abbr.	Common optimization criteria
T	Time (Movement, Wait, Overtime...)
C	Cost (Movement, Wait, Overtime...)
D	Travelled Distance
PP	Patients Preferences
CP	Caregivers Preferences
AS	Accomplished Services

In order to make a detour on the problem encountered in the home health care sector, a classification of recent papers addressing the vehicle routing problem in HHC according to two main criteria. The first one will be the optimization criteria, the second one are the constraints studied. Table I shows the overall objectives encountered in the recent papers and Table II shows the main optimization criteria considered in the home health care issues.

III. PRESENTATION OF THE PROBLEM

The problem dealt with in this article falls under the case of the vehicle routing problem with time windows, synchronization and priority.

Suppose V is the set of requested services and $V1$ the subset of fixed-time requests and $V2$ the subset of flexible hours requests. Each request i is characterized by a location (patient's home) and a time slot $[a_i, b_i]$ where each caregiver (worker) is forced to take and / or collect the sample (blood / urine tube, etc.).

Let e_i be the estimated turnaround time required for sample collection. In the event that a required service has been performed by an external entity, a cost c_i is considered. For specific tests where it is imperative to take / collect a sample, its duration must not exceed a maximum duration between the collection and the performance of the test within the laboratory and is affiliated a variable D_{max_i} , unless the test becomes unnecessary. In general, when the samples are taken, it is necessary that the tubes arrive within a period not exceeding 90 minutes at the laboratory, as for the collection, they are ensured by vehicles managed by the laboratory itself which run continuously in order to collect the tubes across the geographical field where the laboratory performs its services. These critical samples are estimated at 30% of the overall tests carried out by the laboratory in question, which is why the laboratory has delegated this task to a team specially designed for this and which travels only for this kind of case.

To generalize and simplify the problem, we have considered a daily work schedule.

In this article, the main goal is to find a set of possible routes to deliver and collect blood / urine tubes from the laboratory to the homes of patients (clients) and vice versa in an optimal way. For this, we will follow a plan as follows:

- Propose a mathematical formulation of the problem considered.
- Use a linear CPLEX solver to find an exact result of the problem.
- Apply 2 metaheuristic algorithms in order to find an approximate solution to the problem.
- Apply these techniques on a real instance of BioGuich laboratory for a maximum interval of one working day.
- Compare the different results obtained by the techniques applied.

IV. MATHEMATICS FORMULATION

Several mathematical formulations have been proposed in order to modify this type of problem. In this section, we will present a formulation known for its simplicity which was first introduced by [31]. This model is defined by a directed graph where N is the set of sampling requests, D represents the starting point and the end point of each nurse and finally C denotes the laboratory / collection point combination each specified by her collection time in accordance with its transport schedule. Node 0 represents the laboratory and each node represents a patient who needs to be sampled. Each arch has a weight that represents the time the vehicles travel from the laboratory to the patient's home. The time required to charge an entire collection point is denoted by. A set M of nurses is available to slave all requests. Nurses are denoted by the index k such that, and are characterized by a work schedule (for start time and end time), a starting point usually the laboratory and an end point either a collection point or the laboratory. We also introduce the parameter SV which represents a real number and an indicator for each request i which takes the Boolean value 1 if the test is critical and 0 otherwise. In the case of the laboratory whose study is

concerned, the laboratory has only one collection point. This model requires binary decision variables and variables such as:

$$\text{Min } \sum_{i \in N_1} C_i (1 - \sum_{k \in M} \sum_{j \in V/i \neq j} x_{jik}) + (\sum_{k \in M} \sum_{i \in V} \sum_{j \in V/i \neq j} d_{ij} \cdot x_{ijk}) \quad (1)$$

Subject to:

$$\forall i \in N : \sum_{k \in M} \sum_{j \in V/i \neq j} x_{jik} = \sum_{j \in V/i \neq j} x_{ijk} \quad (2)$$

$$\forall i \in V/D, \forall k \in M : \sum_{j \in V/i \neq j} x_{jik} = \sum_{j \in V/i \neq j} x_{ijk} \quad (3)$$

$$\forall k, k' \in M / k \neq k' : \sum_{i \in V/i \neq B_k} x_{iB_k k} = 1 ; \sum_{i \in V/i \neq B_{k'}} x_{iB_{k'} k} = 0 \quad (4)$$

$$\forall k, k' \in M / k \neq k' : \sum_{i \in V/i \neq H_k} x_{H_k i k} = 1 ; \sum_{i \in V/i \neq H_{k'}} x_{H_{k'} i k} = 0 \quad (5)$$

$$\forall k \in M : \sum_{i \in C} x_{H_k i k} = 0 \quad (6)$$

$$\forall i \in N, \forall k \in M \setminus b_i : \sum_{j \in V/i \neq j} x_{ijk} = 0 \quad (7)$$

$$t_i + e_i + d_{ij} \leq t_j + SV. (1 - \sum_{k \in M} x_{ijk}) \quad (8)$$

$$\forall i \in N, \forall j \in C, \forall k \in M : t_i + e_i + d_{ij} \leq t_{jk} + SV. (1 - x_{ijk}) \quad (9)$$

$$\forall i \in C, \forall j \in N, \forall k \in M : t_{ik} + t_{di} + d_{ij} \leq t_j + SV. (1 - x_{ijk}) \quad (10)$$

$$\forall i \in N, \forall k \in M : x_{H_k i k} \cdot (STa_k + d_{H_k i}) \leq t_i \quad (11)$$

$$\forall i \in N, \forall k \in M : t_i + e_i + d_{iB_k} \leq End_k + Ret_k + SV. (1 - x_{ijk}) \quad (12)$$

$$\forall i \in C, \forall k \in M : t_{ik} + t_{di} + d_{iB_k} \leq End_k + Ret_k + SV. (1 - x_{ijk}) \quad (13)$$

$$\forall k \in M : Ret_k \leq Max_{Ret_k} \quad (14)$$

$$\forall i \in N : a_i \cdot \sum_{k \in M} \sum_{j \in V/i \neq j} x_{jik} \leq t_i \leq b_i + Lateness_i \quad (15)$$

$$\forall i \in N : t_i \leq (b_i + Max_{Lateness_i}) \cdot \sum_{k \in M} \sum_{j \in V/i \neq j} x_{jik} \quad (16)$$

$$\forall i \in N / LT_i = 1, \forall j \in C, \forall k \in M : PT_j \cdot drp_{ijk} - t_i \leq DMax_i \quad (17)$$

$$\forall i \in N / LT_i = 1, \forall j \in C, \forall k \in M : t_{jk} \leq PT_j + SV(1 - drp_{ijk}) \quad (18)$$

$$\forall i, j \in V : x_{ijk} \in \begin{cases} 1, & \text{if nurse travel from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases} \quad (19)$$

$$\forall i \in N = N_1 \cup N_2 : t_i : \text{Time of arrival at the patient's home} \quad (20)$$

$$\forall i \in C, \forall k \in M : t_{ik} : \text{Time of arrival of nurse } k \text{ at the collection point with collection time equal to } t_{ik} PT_j \quad (21)$$

$$\forall i \in N / LT_i = 1, \forall j \in C, \forall k \in M : drp_{ijk} \in \begin{cases} 1, & \text{if nurse travel from } i \text{ to } j \\ 0, & \text{sinon} \end{cases} \quad (22)$$

$$\forall k \in M : Ret_k : \text{Nurse's delay } k \text{ (Additional work)} \quad (23)$$

$$\forall i \in N : Lateness_i : \text{Delay of request } I \quad (24)$$

The objective function (1) is to minimize delays while respecting the time windows of the patients and the hourly planning of the nurses responsible for taking and collecting the blood / urine tubes as well as minimizing the total sum of the distances traveled. The constraint (2) indicates that each node can only be visited once and only once. The constraint (3) ensures the continuity of the roads. The constraints (4), (5) and (6) are put in place to define the starting and ending points of each of the routes. Constraint (7) specifies the possible preferences and requirements of certain nurses for a patient. The constraints (8), (9) and (10) represent the duration of a trip between two points. The work schedule is ensured by the constraints (11), (12) and (13) while the constraints (14) and (15) define the time windows of each request established by a patient. The constraint (16) is specially designed to allow delays provided they are delimited. Finally, the constraints (17) and (18) serve to demonstrate that the time elapsed between the collection of a critical test and the moment when it is tested in the laboratory must not exceed a time limit. $DMax_i$. The constraints (19), (20), (21), (22), (23) and (24) define the nature of the decision variables.

V. RESOLUTION APPROACH

This approach is based on the breakdown of the problem into a daily routing problem. First, we will determine the routes for each nurse for a working day. At the beginning, we considered the planning for a large period going from a week to a month, but after a reconsideration and broad reflection we found that this idea is not attractive and that it complicates the problem of planning given the number of constraints to consider is large. To remedy this, it has been found that the best solution is to schedule the nurses' visits in accordance with the demands of the patients day by day and by doing this, a load balancing is achieved.

The experiments we performed using the linear CPLEX solver showed a limitation and that we could only get good results for small instances. The fact that the problem studied in this paper will be applied on a real instance and therefore a practical case, the demands will only increase and for that we are deemed to find another approach that will satisfy the demands of the patients for thus better manage the movements of nurses to the patients concerned, which leads to find a more persistent and more adequate approach such as the approximate methods, namely in this case the metaheuristics of which we will use two methods; the taboo search algorithm and the variable neighborhood search algorithm.

A. Neighborhood Operators

The tabu search and variable neighborhood search algorithms are based on two types of neighborhood operators. The first type is the insertion operator which has three possible movements: - A request which already belongs to a route can be inserted into another route, a request subcontracted by an external entity can be inserted into a route and finally, a request belonging to a route can be removed from a route and added to the set of subcontracting requests.

The second type of neighborhood operator is the permutation operator. This operator has two types of movement. A request belonging to the set of subcontracted requests can be exchanged with a request belonging to a route, the second movement is that a request previously belonging to a route can be exchanged with another request belonging to the same route or to a different route.

B. Algorithms

The tabu search and variable neighborhood search algorithms are based on two types of neighborhood operators. The first type is the insertion operator which has three possible movements: - A request which already belongs to a route can be inserted into another route, a request subcontracted by an external entity can be inserted into a route and finally, a request belonging to a route can be removed from a route and added to the set of subcontracting requests.

The second type of neighborhood operator is the permutation operator. This operator has two types of movement. A request belonging to the set of subcontracted requests can be exchanged with a request belonging to a route, the second movement is that a request previously belonging to a route can be exchanged with another request belonging to the same route or to a different route.

1) *Initial solutions:* The first step to consider before using each of the two methods introduced is to generate a first solution commonly called the initial solution. This solution is built by inserting patient requests into the routes one by one within the best possible position, taking into account their priority. In this case, priority is given to requests with fixed schedules by considering the sizes of the time windows in descending order for each visit of a patient, while the requests with variable schedules, they are processed by the numbers of days remaining until 'at the end of their specified time window. If a request is not inserted, all of these requests will be placed in a list of deferred requests. For critical requests, a collection point is inserted.

2) *The taboo search algorithm:* This algorithm follows a classical structure: from an initial solution, the algorithm moves from the current solution to another solution, exploring the neighborhood while avoiding those which are currently taboo until meeting the stopping condition. In this case, we opted to use the type of insertion operator. The algorithm in question uses a taboo list of recently moved requests. The stop condition used in this case is the maximum number of iterations corresponding to the size of the tabu list.

C. Numerical Results

The laboratory to which we will apply the algorithms and methods described above has provided the study with real instances of a working week, namely a working week starting on Monday and ending on Friday, including the working hours for each stall day. 8:00 a.m. to 6:00 p.m. For the generalization of the studied problem, the staff rest time has not been taken into account and may be the subject of a future work as a perspective. We have generated two groups of instances, the first group is a relatively small and simple set

made up of 10, 15 and 20 requests, with 6, 9 and 12 requests at fixed times respectively. The number of nurses allocated for these requests is usually 1 or 2 and no collection point for this instance group (No critical testing). This small instance is practically tied up for the exact method (MILP) in order to obtain an optimal solution to be able to compare it with the results obtained by the heuristic algorithms. The second group of instances is obtained by the BioGuich laboratory for real requests during a working week, the number of requests for each day generally varies from 150 to 200 in the majority of cases with respectively requests to fixed times from 90 up to 120. We have opted to apply the algorithms proposed for the number of requests of 150, 175 and 200 with respectively fixed time requests of 90, 105 and 120. For critical tests, the percentage is approximately 20% (5 collection points) which will be slaved and will depend on the distance from the laboratory.

Algorithm1. VNS

```

Start
- Initial solution generation ← Solbest
- k ← 1
- As long as a predefined number of iterations without improvement:
- Repeat
- S' ← Shake yourself using the kth operator and start from Solmei
- S'' ← VNS2 starting with S'
- If f(S'') < f(Solbest) then
- Solbest ← S''
- k ← 1
- Else
- k ← k + 1
    
```

The tests of the methods studied in this paper were performed on a 2.4 GHz Intel Core (TM) i7-5500 CPU machine with 8GB of RAM, running Windows 10 Home. The codes were programmed in CPLEX Studio version 12.10 for the mathematical formulation, and Python 3.8 for the heuristic algorithms.

TABLE III. RESULTS FOR SMALL INSTANCES

Instances			MILP	Heuristics	
Patient	Nurse	Collection point	% Optimal	Taboo Research (%)	Variable Neighborhood Search (%)
10	1	0	100	100	80
15	1	0	100	80	60
15	2	0	100	60	40
20	1	0	100	70	50
20	2	0	100	20	30
10	1	1	100	80	90
15	1	1	50	90	70
15	2	1	20	80	20
20	1	1	0	90	90
20	2	1	0	100	100

In order to better understand Table III, each row represents a set of 10 instances. The first three columns represent the size of each instance and is made up of the number of patients, the number of nurses and the collection point. The MILP column represents the percentage of instances where the linear CPLEX solver has found an optimal solution. The maximum execution time has been limited to a maximum of one hour. In the case where CPLEX has not found an optimal solution, we use the best solution found after one hour of execution to compare it with the results of the heuristic algorithms.

The remaining two columns indicate the proportion of instances where each algorithm found a better solution than that found by the linear CPLEX solver. As it is clear, the linear solver always find optimal results than that of heuristics provided that the collection point is not used, but when a collection point comes into effect, the heuristics manage to converge towards a better solution than that of the linear solver. Despite the power of heuristic algorithms compared to the linear solver, this does not prove that the latter have been able to find optimal solutions.

Regarding the speed of the proposed algorithms, it is obvious that the taboo search algorithm is clearly faster than the others, despite this the quality of the solutions produced by the latter are inferior.

After having carried out several tests on small instances, it was deduced that the parameters of the strategic values of the heuristic algorithms are as follows: the size of the tabu list is fixed at 65% of the total number of requests, and the number of iteration without improvement while respecting the best solutions was fixed at 1000. For the variable neighborhood search algorithm, the maximum number of iterations was fixed at 500. These values were chosen in order to be able to obtain results for real instances of the problem investigated in a reasonable time.

Table IV shows the results obtained by the group of instances number 2, of a week of work in the BioGuich laboratory. Each row represents a set of 10 instances. The first three columns show the number of patients, the number of nurses and the number of collection points. Columns 4 and 6 show the total percentage of instances for which effective solutions are found, which can give us an idea of which method gave the best results. Columns 5 and 7 indicate the execution time of each solution for the two algorithms which specify that the variable neighborhood search algorithm is faster than taboo search algorithm while for the quality of the results obtained.

Table V provides details on a work week (End of September 2020) in the BioGuich laboratory with a total of 801 requests executed by 8 nurses during the week.

Table VI and Table VII highlight the objective values of each criterion of the heuristic algorithms studied in this paper.

TABLE IV. RESULTS FOR REAL DATA

Instances			Taboo Search		Variable neighborhood search	
Patient	Nurse	Collection point	% of the best	CPU (ms)	% of the best	CPU (ms)
150	10	5	0	1105	60	59562
150	15	5	40	2560	20	71056
150	20	5	70	3050	10	75487
175	10	5	0	1080	50	68154
175	15	5	0	1700	40	66213
175	20	5	50	4520	30	94054
200	10	5	0	2065	20	78456
200	15	5	20	2310	60	112289
200	20	5	30	7200	20	125394

TABLE V. LABORATORY DATA FOR A WEEK

	Monday	Tuesday	Wednesday	Thursday	Friday
Nb of requests	164	145	158	128	106
Nb of requests at fixed times	128	85	108	94	72
Nb of critical requests	36	24	30	26	18
Number of nurses	8	8	8	8	8

TABLE VI. TABU SEARCH RESULTS

Instances			Taboo Search			
Patient	Nurse	Collection point	Number of outsourced requests	Number of deferred requests	Total delays	Total distance traveled
150	10	5	8	49.3	308.5	1,389.4
150	15	5	4.2	5.8	208.4	2,178.9
150	20	5	2.9	1.2	23.5	1,487.6
175	10	5	18.5	61.6	352.8	1,356.2
175	15	5	3.2	23.8	315.4	2,487.9
175	20	5	4	2.9	81.6	2,154.3
200	10	5	28.4	74.1	295.7	1,348.4
200	15	5	5.2	42.6	367.1	2,259.1
200	20	5	3.5	6.2	236.9	2,478

TABLE VII. VARIABLE NEIGHBORHOOD SEARCH RESULTS

Instances			Variable Neighborhood Search			
Patient	Nurse	Collection point	Number of outsourced requests	Number of deferred requests	Total delays	Total distance traveled
150	10	5	6.4	44.3	306.6	1,176.5
150	15	5	3.2	6.2	216.2	2,126.4
150	20	5	2.9	1.5	26.4	1,619.1
175	10	5	15.3	57.8	364.1	1,247.6
175	15	5	2.5	20.7	302.3	2,392.3
175	20	5	2.8	3.1	85.9	2,276
200	10	5	22.7	72.6	312.2	1,218.8
200	15	5	5.1	38.4	371	2,142.7
200	20	5	3.4	6.9	208.5	2,431.2

VI. CONCLUSION AND PERSPECTIVES

In this paper, we have studied a specific case of the vehicle routing problem for blood/urine samples and collection of samples in the city of Témara located in Morocco. As part of this study, the BioGuich laboratory for medical analyzes since the start of containment in March 2020 due to COVID-19 in Morocco has added the service to travel to patients at their homes in order to take samples and collect required to limit patient movement. For this, it turned out to be necessary for the laboratory to develop a logistics system aimed at optimizing resources for proper management of this service. The result obtained by this study is the development of two meta-heuristics to optimize and plan as best as possible the movements of nurses in order to control the demands of patients, which are constantly increasing day by day. This study was able to save the laboratory considerable time and increase the traceability of the samples collected, which improves the quality of the tests carried out. For future work, it is necessary to add human constraints, such as breaks for nurses, introduce a penalty system for late tests, refine the available routes in order to minimize travel times. This study was able to save the laboratory considerable time and increase the traceability of the samples collected, which improves the quality of the tests carried out.

REFERENCES

[1] N. Velasco, P. Dejax, C. Guéret & C. Prins (2012) A non-dominated sorting genetic algorithm for a bi-objective pick-up and delivery problem, *Engineering Optimization*, 44:3, 305325, DOI: 10.1080/0305215X.2011.639368

[2] Parragh et al., 2008a S.N. Parragh, K.F. Doerner, R.F. Hartl A survey on pickup and delivery problems. Part I: transportation between customers and depot *Journal für Betriebswirtschaft*, 58 (2008), pp. 21-51

[3] Parragh et al., 2008b S.N. Parragh, K.F. Doerner, R.F. Hartl A survey on pickup and delivery problems. Part II: transportation between pickup and delivery locations *Journal für Betriebswirtschaft*, 58 (2008), pp. 81-117

[4] Fahle, T : Production and transportation planning modeling report. Report, University of Paderborn (2001).

[5] Begur, SV, Miller, DM, Weaver, JR: An integrated spatial decision support system for scheduling and routing home health care nurses. *Interfaces* 27, 35-48 (1997).

[6] Akjiritakarl, C., Yenradee, P., & Drake, PR 2007. PSO-based algorithm for home care worker scheduling in the UK. *Computers & Industrial Engineering*, 53 (4), 559-583.

[7] Rousseau, LM, Gendreau, M., & Pesant, G. 2013. The Synchronized Dynamic Vehicle Dispatching Problem. *INFOR: Information Systems and Operational Research*, 51 (2), 76-83.

[8] Liu, R., Xie, X., Augusto, V., & Rodriguez, C. (2013). Heuristic algorithms for a vehicle routing problem with simultaneous delivery and pickup and time windows in home health care. *European Journal of Operational Research*, 230 (3), 475-486.

[9] Ettazi, H, Rafalia, N, Abouchabaka, J: GRASP for the Vehicle Routing Problem with Time Windows, Precedence, Synchronization and Lunch Break constraints. *Proceedings of the 36rd International Business Information Management Association Conference, IBIMA 2020 (InPress)*.

[10] Ettazi, H, Rafalia, N, Abouchabaka, J: A Metaheuristics methods for The VRP in Home Health Care by minimizing fuel consumption for environmental gain E3S Web Conf., 234 (2021) 00094 DOI: <https://doi.org/10.1051/e3sconf/202123400094>

[11] Cheng, E, & Rich, J.L. 1998. A home health care routing and scheduling problem. *Technical report CAAM TR98-04, Rice University*.

[12] Eveborn, P., Flisberg, P., & Rönnqvist, M. 2006. Laps Care – an operational system for staff planning of home care. *European Journal of Operational Research*, 171(3), 962-976.

[13] Bertels, S., & Stefan, T. 2006. A hybrid setup for a hybrid scenario : combining heuristics for the home health care problem. *Computers & Operations Research*. 33(10), 2866-2890.

[14] Doerner, K., Focke, A. & Gutjahr, W.J. 2007. Multicriteria tour planning for mobile healthcare facilities in a developing country. *European Journal of Operational Research*, 179(3), 1078-1096.

[15] Bredström, D., & Rönnqvist, M. 2007. A branch and price algorithm for the combined vehicle routing and scheduling problem with synchronization constraints. *NHH Dept. of Finance Management Science Discussion Paper*.

[16] Bredström, D., & Rönnqvist, M. 2008. Combined vehicle routing and scheduling with temporal precedence and synchronization constraints. *European Journal of Operational Research*, 191(1), 19-31.

[17] Bräysi, O., Dullaert, W., & Nakari, P. 2009. The potential of optimization in communal routing problems : case studies from finland. *Journal of Transport Geography*, 17(6), 484-490.

[18] Trautsamwieser, A., & Hirsch, P. 2011. Optimization of daily scheduling for home health care services. *Journal of Applied Operational Research*, 3(3), 124-136.

[19] Redjem, R., & Marcon, E. 2015. Operations management in the home health care services : a heuristic for the caregivers' routing problem. *Flexible Services and Manufacturing Journal*, 1-24.

[20] M.S Rasmussen, T. Justesen, A. Dohn and J. Larsen. 2012. The Home Care Crew Scheduling Problem: Preference-Based Visit Clustering and Temporal Dependencies. *European Journal of Operational Research* (3), p. 598-610.

[21] Coppi, A., Detti, P., & Rafaelli, J. 2013. A planning and routing model for patient transportation in health care. *Electronic Notes in Discrete Mathematics*, 41, 125-132.

[22] Afifi S., Dang, D.C., & Moukrim, A. 2013. A simulated annealing algorithm for the vehicle routing problem with time windows and synchronization constraints. Pages 259-265: *Learning and Intelligent Optimization*. Springer.

[23] Labadie, N., Prins, C., & Yang, Y. 2014. Iterated local search for a vehicle routing problem with synchronization constraints. Pages 257-263 of : *ICORES 2014-Proceedings of the 3rd International Conference on Operations Research and Enterprise Systems, Angers, Loire Valley, France*.

[24] Ceselli, A., Righini, G., & Tresholdi, E. 2014. Combined location and routing problems for the drug distribution. *Discrete Applied Mathematics*, 165, 130-145.

- [25] Redjem, R., Kharraja, S., Xie, X., & Marcon, E. 2012 ; Routing and scheduling of caregivers in home health care with synchronized visits. In : 9th International Conference on Modeling, Optimization & SIMulation.
- [26] Issaoui, B., Zidi, I., Marcon, E., & Ghedira, K. 2015. New Multi-Objective Approach for the Home Care Service Problem Based on Scheduling Algorithms and Variable Neighborhood Descent. *Electronic Notes in Discrete Mathematics*, 47, 181-188.
- [27] Ait Haddadene, S.R, Labadie, N., & Prodhon, C. 2016. A GRASP \times ILS for the vehicle routing problem with time windows, synchronization and precedence constraints. *Expert Syst. Appl*, Vol. 66, 274-294.
- [28] Shahnejat-Bushehri, S., Tavakkolo-Moghaddam, R., & Momen, S., Ghasemkhani, A., Tavakkolo-Moghaddam, H. 2019. Home Health Care Routing and Scheduling Problem Considering Temporal Dependencies and Perishability with Simultaneous Pickup and Delivery. *IFAC-PapersOnline*, 53(13), 118-123.
- [29] Borchani, R., Masmoudi, M., & Jarbaoui, B. 2019. Hybrid Genetic Algorithm for Home Health care routing and scheduling problem. *CoDIT*, 1900-1904.
- [30] Bazirha, M., Kadrani, A., & Benmansour, R. 2020. Daily Scheduling and Routing of Home Health Care with Multiple Availability Periods of Patients. *Variable Neighborhood Search*, 178-193.
- [31] Fisher, ML, & Jaikumar, R. 1981. A generalized assignment heuristic for vehicle routing. *Networks*, 11 (2), 109-124.

A Computer-assisted Collaborative Reading Model to Improve Reading Fluency of EFL Learners in Continuous Learning Programs in Saudi Universities

Abdulfattah Omar¹, Mohamed Saad Mahmoud Hussein², Fahd Shehail Alalwi³
Department of English, College of Sciences and Humanities¹
Prince Sattam Bin Abdulaziz University, Faculty of Arts, Port Said University, Egypt¹
Faculty of Education, Assiut University, Egypt²
Department of English, Prince Sattam Bin Abdulaziz University, Saudi Arabia³

Abstract—Reading is not synonymous to comprehension; rather it is a prerequisite that doesn't, by itself, guarantee comprehension. This is to say that being efficient in decoding letters, syllables and whole words and recognizing vocabulary does not ensure natural r automatic comprehension. Fluency seems to be the bridge between the mastery of the mechanics of reading and the dynamics of comprehension. Abundant research exists that explores how to improve reading skills of EFL learners at Saudi universities. However little, if any, of this array of research sought to discern the potential effects of educational technology on the fluency of struggling readers in continuous learning programs. To fill this gap, this study seeks to probe the multi dimensions of the problem and suggest ways to solve it. For this purpose, 24 EFL lecturers from three Saudi universities were selected and interviewed. A suggested computer-assisted collaborative reading model was put forth to be applied in the three universities. Students were diagnosed by their instructors as gaining relatively enough grasp of decoding skills at the multi-levels of orthographic knowledge, mono and polysyllabic words, but exhibit slow and inaccurate reading indicating reprehensive symptoms for a fluency problem. The lecturers explained that the disappointment resulting from learners' inability to reach comprehension despite mastering decoding skills influences their attitudes towards reading and language learning, bringing about reading apathy and low self-esteem. The proposed model is designed to enhance reading fluency which is perceived as the underlying problem that makes the reader struggle. It is to be delivered partly individually and partly collaboratively online. Collaboration also is operated via face to face instruction especially in teaching the reading strategy. In doing so, the procedures followed are in line with the blended learning. The findings indicate clearly that the proposed model was successfully used to improve reading fluency through accelerating the different reading subskills for decoding and create positive attitudes toward reading. The results highlight the importance of establishing a level of automaticity that gives rise to the higher skills of comprehension.

Keywords—Collaborative reading model; computer-assisted language learning (CALL); computer-based instruction; EFL learners; fluency; Saudi Universities; struggling readers

I. INTRODUCTION

English is gaining prominence in the Saudi community .This is reflected through the pivotal position the educational system assigns to teaching English. This is due to the role

English plays as a lingua franca and the different significant social, cultural and economic functions it fulfills along with being the international language of science and technology. The increasing demands for more proficient language learner locally and overseas impose pressure on educational stakeholders to secure such employable working force. Therefore, it is becoming integral part in all university programs [1, 2].

In response to such demands, a wide range of studies have been conducted to explore the possible factors, cognitive and affective, that contribute to the enhancement of the didactic practice and the development of the language skills in general and the reading skills in particular [3-6]. Nevertheless, the unique profile typical to continuous learning students compared to standard programs in terms of the program's admission prerequisites entails special attention to be directed to investigating the teaching and learning practices of such sector especially concerning struggling readers.

Reading, despite its undeniable significance for success in life and academia, is a problematic skill that a large number of continuous learning students stumble about. The cumulative struggling with reading is attributed to students' failing to master reading fluency. Therefore, a computer assisted collaborative reading model is suggested to enhance the different subskills making up fluency. The study attempts to answer the following questions:

- What are the missing reading skills that interfere with comprehension given the mastery of decoding skills?
- To what extent the computer assisted collaborative reading model is able to address the missing component?

A case study design was adopted to answer the aforementioned questions. Twenty four EFL lecturers and faculty members from three Saudi universities were selected and interviewed about the characteristic profile of the struggling students who flounder with reading in spite of the fact that they are good at decoding with its different subskills. A model was designed accordingly and the selected lecturers and faculty members were assigned to guide the collaborative part of the model and observe students' behavior and progress. The selected participants were then interviewed about the

usefulness of the proposed model in addressing the problems struggling learners face in these programs and mitigating the affective consequences arising as by-products of the problem.

The remainder of the article is organized as follows. Section 2 is a theoretical framework. It defines the key concepts of the study and sets the relation between computer-assisted collaborative models and reading fluency. Section 3 is a brief survey of the previous studies on the integration of CALL systems in EFL reading in the Saudi contexts. Section 4 is Methodology. It outlines data collection and analysis procedures. Section 5 reports on the results of the study. This is a qualitative data analysis of the data derived from the interviews with the selected EFL lecturers. Section 6 is Conclusion. It summarizes the main findings of the study and discusses the implications of the study to teaching English in continuous programs in the Saudi universities and to further research.

II. THEORETICAL BACKGROUND

With the wide proliferation of technology and its powerful and reconfiguring penetration into most of our life domains including banking and shopping, the integration of digital technologies into teaching and learning language has become inevitable. This incorporation has resulted in changes within the structure of the teaching/learning environment away from the longstanding teacher dominated pedagogy towards a more student-centered leaning situation [7, 8].

The integration of technology into learning contexts has brought about new patterns concerning the role of both the teacher and the learner. This necessitates that both teachers and learners adapt to the new roles; the teacher has to be best prepared to step a little back and be sufficed with the facilitator role and the learners should develop a sense of responsibility to be up to the more autonomy granted to them [9]. Educational technologies afford learners broader opportunities through a flexible learning environment that respects their needs and ability levels and keeps up with their pace of learning. In this regard, e-learning systems extend the traditional learning space borders beyond the narrow sense of the regular classroom creating new and more promising scenery for teaching.

In a similar vein, Stanley and Thornbury [10] argue that technology is increasingly becoming integral part of learning delivery and that it opens the door wide to more innovative teaching practices. The introduction of technological advancements as instructional tools to teach language, which is termed as Technology Enhanced Language Learning (TELL), has contributed to addressing many of the problems learners used to encounter in their language classrooms. Under the broad term of (TELL), fall other disciplines including computer assisted language learning (CALL), mobile Assisted language learning (MALL) and Computer Mediated Communication (CMC). The research into how technology can be applied to language teaching has given rise to the emergence of some now widely touted fields like MALL and CALL [11, 12].

CALL is a learning approach that makes use of the computer software and the internet-based resources as a means

to deliver, reinforce and assess learning [13]. The advent of Web 2.0 tools with its characteristic features of interactivity and other aspects that facilitate collaborative learning provided the practical grounds for utilizing computers for teaching and learning purposes [14]. In its integration of education and technology, CALL comprises most ICT applications and learning approaches relevant to L2 acquisition [15].

With the unprecedented development of social media systems, networks, and platforms in recent years, social media as ICT tools have been extensively used instructionally [16, 17]. Research on social media used as pedagogical vehicles has yielded positive results. For example, Wang and Vasquez [18] reported that Facebook was effective in improving Chinese language learners' performance in writing. Similarly, Hamat and Hassan [19] surveyed Malaysian university undergraduates' perceptions about how helpful social networking services are in learning English and in what areas they have the most effect. Findings revealed that (99.7%) use the sites for learning English and the areas which benefited most were reading, writing, vocabulary acquisition and communication. In his investigation of the Bangladeshi university EFL learners' perceptions about how effective social media are in improving speaking skill, Mitu [20] reported an affirmative positive relationship. Digital games, another manifestation of CALL, have rendered promising results in terms of their ability to "motivate students, increase time on task and encourage collaboration and situated communication" [9].

There is a concern, among some scholars, that learning via computers can be a socially isolating experience due to the physical separation between students. This is to mean that collaboration should be an integral part of CALL to mitigate this concern. Most importantly, teamwork seems to be an inherent human desire and perhaps a necessity. This applies to learners who mostly like to work and learn together [21]. Collaborative learning is essentially and organically related to e-learning where students work together to achieve a common task via different ICT tools and devices including live chat or instant messages in order to expand their knowledge and make use of one another's strengths [22]. In fact, collaboration is significantly useful in language acquisition as it serves as a scaffolding vehicle through which learners mutually help each other enhance their language ability and facilitate meaning communication. It also enhances social interaction and hones thinking skills. Moreover, it familiarizes learners with the model practice expected both in the academia and beyond [21].

CALL, MALL and TELL are essentially about the integration of technology in second or foreign language acquisition. Kasemsap [23] argues that these three environments share the powerful potentials of improving the motivation and attitudes of language learners toward language learning. It can be argued that the use of CALL and MALL is essentially the same given that applications accessed by mobile can also be accessed by the computer. According to Egbert [24], CALL does not necessarily involve computers as he argues that CALL means learners' learning language in any context with, through, and around computer technologies. It is

also possible to take the computer technologies a step further to include digital technologies. This might be the reason why Ogata, et al. [25] consider MALL as falling under CALL as they use the term “computer assisted mobile learning.

Since most practical utilizations of both CALL and MALL are essentially based on the interactive feature associated with the emergence of Web 2.0 tools, it seems valid to claim that both the two systems differ only in the degree of mobility not in purpose of use or the underlying rationale. Yaman and Ekmeççi [26] consider the shift as just a transfer of functions from the computer to the small mobile device. Kukulska-Hulme and Shield [27] contend that the size and mobility aspects are the distinctive features of MALL that afford learners new ways of learning, emphasizing continuity or spontaneity of access across different contexts of use”. They think that the two terms are complementary and each system has its relevant conditions that make its use more advisable than the other. Thus, occasions determines which system is used.

Fluency is simply defined as reading accurately and quickly in a conversational like manner. Thus, for fluency to embryonate and materialize, these three elements should occur together [28-30]. Fluency is a complex multilayered construct that deals with many levels of processing including “decoding fluency, processing speed, vocabulary, letter sound fluency, and sight word fluency” [28]. The complicated and stratified nature of fluency is clearly reflected in the definition given by Fuchs, et al. [31] which states that fluency is made up of a group of sub-processes including a reader’s perceptual skill at automatically translating letters into coherent sound representations, unitizing those sound components into recognizable wholes, and automatically accessing lexical representations, processing meaningful connections within and between sentences, relating text meaning to prior information, and making inferences to supply missing information [31].

This is to imply that reading fluency deals with the lower and the higher order reading processes constituting a link between them through a dynamic reciprocal backward and forward relationship [28]. Decoding and comprehension remain latent until fluency connects the two extremes of the circuit. Samuels [32] puts this in other words as he considers fluency as “decoding and comprehending at the same time”. For him accuracy, speed and prosody are operational of the construct. Hence, fluency can be conceptualized as the automatic utilization of all reading processes and sub-processes, a conceptualization advocated by Wolf and Katzir-Cohen who argue that “fluency is influenced by the development of rapid rates of processing in all the components of reading [33].

It is the conscious deliberate analysis of each or any of the different decoding levels that slow down processing and might get the reader to be stuck at one or more of the levels. This means that automaticity must be at every level and there must be gracious and smooth transfer between levels [28]. The automaticity as the cornerstone of fluency is underpinned on the theory of automatic information processing in reading suggested by LaBerge and Samuels [34]. It posits that the processes and sub-processes of reading decoding should be

executed automatically with minimal cognitive effort so that the saved mental capacity is directed to the more important goal of reading, which is making sense of the read material. According to the theory, poor reading is likely to occur when the readers use up most of their cognitive capacity in the lower surface-level processes of reading leaving little mental resources for comprehension. Reaching automaticity goes through some steps; a teacher is needed to instruct learners in developing the skills at the conscious level, as a first step. Then comes the role of the repeated practice that takes the performance to the sphere of automatic practice [35].

III. LITERATURE REVIEW

The Saudi research community in the applied linguistics arena is quite active as for exploring the multifaceted aspects of language teaching and learning. This is also true to the research efforts that addressed the digital learning applications to the development of the reading comprehension skill. This body of research, with its breadth and width, has especially highlighted the unique usefulness of CALL systems in leveraging reading comprehension among the EFL university learners in Saudi Arabia [36-40].

Learning approached via technological devices is privileged by many advantageous aspects which traditional learning falls short of reaching. Bensalem [41] argues that manipulating the CALL rationale to teach reading pays off positive results since these systems facilitate the interaction with the texts, provide for the different learning styles and meet the differentiation considerations necessary to reach all learners. The inherent engagement friendly nature of such systems and the provision for individualized learning and tailoring it to suit the needs and ability level of the learners qualify it as an effective tool for addressing the non-intelligence affective factors such as shyness and poor attitudes. This promising influence of CALL is extended to all levels of sub-skills constituting the overall reading skill. Hassan Taj, et al. [42], for example, reported that CALL is instrumental in improving many of the reading related skills including decoding, word recognition and retention in addition to enhancing the working memory.

Abanomey [43] investigated the influence of the internet reading on the overall reading performance of the undergraduate students of Riyadh College of Technology. Results showed that internet-based reading has more positive effect on reading than the print-based reading. He attributed the results to the motivation inspired by using the internet and to reported that the internet reading enables students to select the proper skill and strategies as needed.

Likewise, wide range of research has covered the scope of the mobile technology in education. Alshammari, et al. [44] surveyed the university tertiary students’ attitudes towards the uses of WhatsApp to conclude that students hold positive attitudes towards such mobile text messaging applications. Moreover, Khojah and Thomas [39] explored the potential effect of the MALL activities on the students in reading classes. Results disclosed significant increase in students’ reading achievement, attention, participation, and volunteering. High motivation and positive reading attitudes were also reported. Keezhatta and Omar [1] explored the

potential relation between MALL and reading comprehension and finding indicated a significant effect of MALL on reading due to the motivating learning environment the system creates. Albiladi and Alshareef [45] surveyed the Saudi English teachers' perceptions about tablets' incorporation in language teaching. It turned out from results that teachers think tablets have facilitating instructional benefits.

Research has also pointed out that CALL has positive effect on the affective side of the learners including motivation, attitude and self-esteem. In their study done on Saudi EFL secondary students, Keezhatta and Omar [1] found that digital technologies positively impacted students' motivation and attitudes toward language learning. They explained that the effect is due to the motivation triggering nature inherent in the instructional computer setting and the reciprocal relation between reading and motivation. The digital technologies context afforded students with some important affective weapons; it equipped them with a sense of security that helped them to build confidence in their ability; a sense of control and as thus responsibility for their learning. Likewise, Alotaibi [46] in his study conducted on university undergraduates reiterated similar results concerning the positive effect of CALL on learners' motivation. Alotaibi [46] documented a strong relation between mobile-assisted language learning (MALL) and enhancing reading motivation.

Beyond the Saudi EFL setting but still within the broader EFL context, technology-mediated instruction proved effective in teaching reading. Varol and Erçetin [47] compared the use and non-use of glosses and hyperlinks at the lexical and topic levels and found out that this technological intervention helps significantly with word recognition, but that effect was not observed for reading comprehension. Alharbi [36] conducted a study on Saudi undergraduate EFL students in Qassim University and came to similar results concerning relation between reading and glossing. Similarly, Ali [48] compared a computer based instruction with the teacher based instruction as for their effect on three reading specific skills, namely speed, vocabulary acquisition and comprehension. Participants were undergraduate students from two universities based in Oman and UAE. Results were in favor of the computer assisted reading. Three considerations were suggested to account for the findings: the engaging activities that capture learner's attention and motivate them to read; the immediate feedback and the sense of autonomy and control learners felt.

Mahmoudi, et al. [49] investigated the attitudes of Iranian postgraduate students in a university in Malaysia toward computer assisted English language learning (CAELL) and their performance concerning the English language vocabulary. Results revealed that learners have positive attitudes to CAELL and a mutually positive correlation was established between the attitude and performance in vocabulary. Digital games, another manifestation of CALL, has rendered promising results/in terms of their ability to motivate students, increase time on task and encourage collaboration and situated communication [9].

Enriching the electronic environment with the social aspects of learning is also highlighted by research in the EFL

context. Lan, et al. [50] used MALL environment as a mode for a peer-based collaborative learning to examine the effect of this mobile-device-supported peer-assisted learning (MPAL) on primary school's reading performance. The intervention was found out to have improved both collaboration level and the reading motivation. In the same fashion, Chen, et al. [51] used the web annotation system as a web-based collaborative reading annotation system (WCRAS) together with gamification mechanisms to promote the reading performance of Taiwan elementary schools and positive results were reported. Yang, et al. [52] investigated the effect of the synchronous form of collaborative learning employing Group scribble on the reading comprehension of Chinese primary schools. Results showed better reading comprehension and enhanced motivation to learn collaboratively through Group scribble. Ae-Hwa, et al. [53] examined the effect of Computer-Assisted Collaborative Strategic Reading (CACSR) program on disabled students' reading comprehension performance. Findings revealed that the program has significant effect on reading.

To summarize, digital technologies with their different demonstrations, including CALL, constitute a pivotal pedagogical vehicle in foreign language teaching and learning. Such educational technologies are especially useful in teaching reading comprehension due to their flexible nature and their potentials in providing individualized reading instruction that meets the varied needs of the learners and keep up with their pacing. In the Saudi context, prolific research has been done on the different aspect of reading comprehension. However, little is devoted to address the struggling readers' profile of the continuous learning programs in Saudi universities, even less is directed to the reading fluency. Hence, this study attempts to fill this gap through targeting this population with a computer-assisted collaborative reading model intended to improve their overall reading performance with special emphasis on reading fluency as a mediating reading process necessary to achieve comprehension.

IV. METHODS AND PROCEDURES

Both male and female lecturers were selected. There were 13 female lecturers and 11 male lecturers. It should be noted that sex segregation is still imposed in the Saudi universities, even with the drastic social and political developments and changes the country is witnessing today. In this regard, female lecturers are more likely to teach in female sections only. Although it is not a rule, female lecturers are not normally selected for teaching male students even in distance and online programs.

This study is based on a case-study design. It is limited to the continuous education programs in the Saudi universities. In so doing, the tool of unstructured interviews was adopted. Interviews took place between September-December, 2020 in three Saudi universities: King Abdulaziz University, King Saud University, and Prince Sattam Bin Abdulaziz University. Twenty four EFL lecturers in three Saudi universities were selected. Only EFL lecturers with first-hand experience in teaching in continuous learning programs in the Saudi universities were included in the study. Representativeness

was also considered for reliability and generalizability purposes. The participants come from different backgrounds.

The participants represent different age groups. They also occupy different positions (language instructors, lecturers, and faculty members). They also come from different countries including Egypt, Jordan, Morocco, Saudi Arabia, and Tunisia.

For data analysis purposes, thematic content analysis was adopted. The rationale is that thematic content analysis can be usefully used to address many of the limitations including selectivity and subjectivity that are always associated with qualitative data analysis approaches [54-58].

The interviews were conducted at two subsequent rounds. First, the participants were asked about the challenges and needs of the EFL learners in the continuous learning programs. Based on these discussions, a collaborative reading model through in-session courses was developed. Second, they were asked about the effectiveness and usefulness of the proposed model.

Thematic content analysis is now supported with different computational systems whose function is to derive the important and relevant information within the qualitative data sets [59]. These computational systems are designed to help researchers with preparing, coding, and analyzing data [60]. For the purposes of the study, ATLAS.ti was used. This is a qualitative data analysis software developed by Scientific Software Development Company in 1993 to address the needs of academic and industry researchers who are involved in qualitative data analysis [61].

ATLAS.ti was selected for convenience reasons. It supports qualitative analysis of large bodies of textual, graphical, audio and video data. It can be usefully used to arrange, reassemble, and manage materials in systematic ways. Furthermore, it has been proved effective in education, teaching, and applied linguistic studies [62-64]. The use of ATLAS.ti includes six main steps or stages. These are pre-processing of data, input data into the computer using ATLAS.ti, exploration of the material, coding, categorization, and treatment of the results [65]. These are shown in Fig. 1.

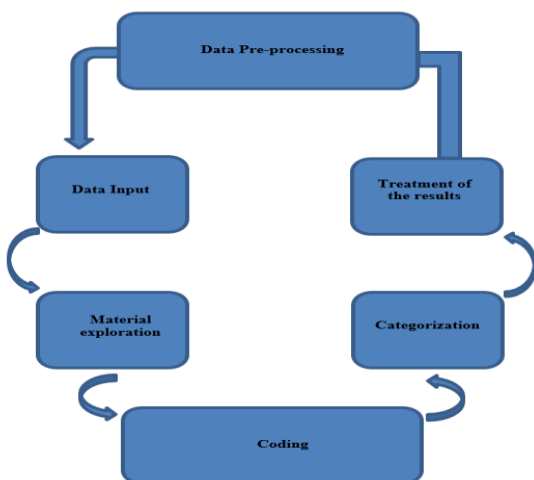


Fig. 1. Steps of Performing Thematic Content Analysis using ATLAS.ti.

All the data were pre-processed, entered into the software, explored, and coded. They were then categorized and classified into the common and key themes mentioned by the participants. One advantage with the software processing is that it reduces selectivity and subjectivity which are common problems with qualitative data analysis. Based on data processing procedures, the participants focused on some common themes including the unique nature of EFL learners in the continuous learning programs, reading challenges, and the special needs of the EFL learners in the continuous learning programs, and learners' performance before and after receiving the in-session course. These are discussed in detail in the next section.

V. ANALYSIS AND DISCUSSIONS

The participants almost agreed that the EFL learners in the continuous learning programs in the Saudi universities face unique linguistic challenges due to different reasons including the Saudi educational system itself which used to focus on Muslim teachings where English was used by some as a threat to the Arab and Muslim identity [66]. These perceptions had negative implications to teaching English as a second language in Saudi Arabia for decades. For many years, English was not taught at primary schools lest students are badly influenced by the English structures [1]. To a large extent, English language teaching was very limited in the Saudi schools and universities [67].

The findings also reflect clearly that reading is the most challenging language skill to the EFL learners in the continuous learning programs. This is clearly reflected in the students' scores. According to the participants, students usually score the less in reading courses and activities. They attributed the failure of the students to some reasons including the differences between Arabic and English, lack of prerequisite linguistic knowledge, study materials, and enrollment conditions. The interviewees pointed out that though students received instruction in decoding at the different levels, they often get stuck at the decoding level failing to make sense of what they read. Such recurrent failure get them disappointed demotivated to read more due to their low self-esteem and thus perpetuating the problem.

Given the different nature of the continuous learning programs, the participants almost agree that the linguistic challenges and needs of the EFL learners in these programs are in many ways different from those of their counterparts in standard programs. Struggling readers in the continuous learning programs need a learning environment that considers their social and professional backgrounds. Many of these learners are full-time employees and the vast majority of them have not been exposed to reading English texts for years.

In light of these findings, it can be claimed that such weakness profile is indicative of a reading fluency problem and accordingly a computer assisted collaborative reading model was suggested to enhance automaticity of the different reading sub-skills with the purpose that such intervention could bring about the desired comprehension and address the problematic affective byproducts of demotivation, negative attitudes and low self-esteem. The rationale is that the integration of CALL systems have been usefully used over the

recent decades in improving different reading skills of EFL learners including word recognition and decoding skills within a short period of time. Furthermore, CALL systems and activities are usually associated with creating motivating learning environments. According to Hubbard [68], CALL systems have the potentials of developing L2 reading skills in a limited span of time compared to traditional teacher-led models. They are also found to be a source of great excitement and enjoyable. Furthermore, numerous studies reported the effectiveness of CALL systems in relation to adult learners with little linguistic background [69, 70].

The proposed computer assisted collaborative reading model was voluntarily conducted by the participant EFL lecturers and faculty members through in-session courses to help struggling readers with the reading courses and improve their reading performance through focus on practice of the different reading skills so as to reach the level of automaticity and the employment of reading strategies in a collaborative mode.

The interviewees finally stressed that the learners responded positively to the proposed model which had a significant impact on the learners' performance and overall achievement. They outlined that the proposed model provided the learners with a self-paced learning environment in which they will work with high level of interest at a faster pace. They added that the model improved the learners' skills in pronunciation, vocabulary, use of words in different contexts, and comprehension. They also reported that learners' skills in relation to vocabulary recognition and text understanding were considerably improved.

Results indicate that the model proved effective in improving students' reading fluency and reading comprehension performance. In addition, the affective profile of the students improved as a result. This can be attributed to some factors; dealing comprehensively with the reading process from its two extremes, decoding and comprehension, through accelerating decoding pace to the level of automaticity and the application of strategic reading to the level of comprehension; the collaborative mode of the computer assisted language learning was useful in many ways including securing the social interaction and scaffolding students into more reflection on their mental processing through comparing their performance with their peers.

It can be finally suggested that there is a necessity for adding the social aspects of learning to the computer-based learning and to get both teachers and students prepared for such shift from class based to virtual space-based learning. In addition, the face-to-face interaction between students and their teacher is necessary which underscores the significance of the blended learning and stress the fact that technology will not replace teachers; teachers who use technology will replace those who don't.

VI. CONCLUSION

The study attempted to discern why decoding supplementary intervention do not make a good reader; they produce symbol decoder at some degree; yet they fail to qualify students into comprehenders. Interviews were done

with instructors selected from three universities to draw out the essential information underlying such reading deficiency. Interviewees pointed out that though students received instruction in decoding at the different levels, they often get stuck at the decoding level failing to make sense of what they read. Such recurrent failure gets them disappointed and demotivated to read more due to their low self-esteem and as thus perpetuating the problem. Based on the interview results, researchers supposed that such weakness profile is indicative of a reading fluency problem and accordingly a computer assisted collaborative reading model is suggested to enhance automaticity of the different reading sub-skills with the aim that such intervention could bring about the desired comprehension and address the problematic affective byproducts of demotivation, negative attitudes and low self-esteem. It was obvious that the proposed model was effective in improving students' reading fluency and comprehension. The students' affective profiles changed as a result. Though the study is limited to addressing the reading challenges and needs of EFL learners in the continuous learning programs in the Saudi universities, the findings of the study can be applicable to other populations.

ACKNOWLEDGMENT

We take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Deanship of Scientific Research, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

REFERENCES

- [1] M. S. Keezhatta and A. Omar, "Enhancing Reading Skills for Saudi Secondary School Students through Mobile Assisted Language Learning (MALL): An Experimental Study," *International Journal of English Linguistics*, vol. 9, no. 1, pp. 437-447, 2019.
- [2] A. Hamdan, *Teaching and Learning in Saudi Arabia: Perspectives from Higher Education*. New York: Springer, 2015.
- [3] A. Al Nooh, "The Effectiveness of Reading Techniques Used in a Saudi Arabian Secondary School Classroom as Perceived by Students and Teachers: A Study of Methods Used in Teaching English and their Effectiveness," *Arab World English Journal*, vol. 4, no. 3, pp. 331- 345, 2013.
- [4] A. Alqahtani, "Why Do Saudi EFL Readers Exhibit Poor Reading Abilities?," *Journal of English Language and Literature*, vol. 6, no. 1, pp. 1-15, 2016.
- [5] K. Al-Nafisah and R. d. A. Al-Shorman, "Saudi EFL students' reading interests," *Journal of King Saud University - Languages and Translation*, vol. 23, no. 1, pp. 1-9, 2011/01/01/ 2011.
- [6] C. Moskovsky and M. Picard, *English as a Foreign Language in Saudi Arabia: New Insights into Teaching and Learning English*. London; New York: Routledge, 2019.
- [7] M. Weimer, *Learner-Centered Teaching: Five Key Changes to Practice*. Wiley, 2008.
- [8] W. Ng, *New Digital Technology in Education: Conceptualizing Professional Learning for Educators*. Springer International Publishing, 2015.
- [9] M. Thomas, H. Reinders, and M. Warschauer, *Contemporary Computer-Assisted Language Learning*. Bloomsbury Publishing, 2012.
- [10] G. Stanley and S. Thornbury, *Language Learning with Technology: Ideas for Integrating Technology in the Classroom*. Cambridge: Cambridge University Press, 2013.
- [11] M. Thomas and C. Schneider, *Language Teaching with Video-Based Technologies: Creativity and CALL Teacher Education*. London; New York: Routledge, 2020.

- [12] B. Zou and M. Thomas, *Recent Developments in Technology-Enhanced and Computer-Assisted Language Learning*. Hershey, Pennsylvania: IGI Global, 2019.
- [13] G. Davies and J. Higgins, *Computers, Language and Language Learning*. Centre for Information on Language Teaching, 1983.
- [14] M. Warschauer and D. Healey, "Computers and language learning: an overview," *Language Teaching*, vol. 31, no. 2, pp. 57-71, 2009.
- [15] C. A. Chapelle, *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press, 2001.
- [16] W. M. Chan, K. N. Chin, M. Nagami, and T. Suthiwan, *Media in Foreign Language Teaching and Learning*. De Gruyter, 2011.
- [17] K. Seo, *Using Social Media Effectively in the Classroom: Blogs, Wikis, Twitter, and More*. London; New York: Routledge, 2012.
- [18] S. Wang and C. Vasquez, "Web 2.0 and second language learning: What does the research tell us?," *CALICO Journal*, vol. 29, no. 3, pp. 412-430, 2012.
- [19] A. Hamat and H. A. Hassan, "Use of Social Media for Informal Language Learning by Malaysian University Students," *3L: The Southeast Asian Journal of English Language Studies*, vol. 25, no. 4, pp. 68-83, 2019.
- [20] F. Mitu, "Mitu (2020) English Language Teaching Materials and Testing Methods Nexus at the Primary Level Education in Bangladesh," *BUBT Journal*, vol. 11, pp. 1-18, 2020.
- [21] K. Beatty, *Teaching & Researching: Computer-Assisted Language Learning*. London; New York: Routledge, 2013.
- [22] S. G. McCafferty, *Cooperative learning and second language teaching*. Cambridge: Cambridge University Press, 2006.
- [23] K. Kasemsap, "Foreign Language Learning: CALL, MALL, and Social Media Perspectives," in *Handbook of Research on Technology-Centric Strategies for Higher Education Administration Advances in Educational Marketing, Administration, and Leadership* Hershey, Pennsylvania: IGI Global 2017, pp. 137-158.
- [24] J. Egbert, *CALL Essentials: Principles and Practices in CALL Classrooms*. TESOL Publications, 2005.
- [25] H. Ogata, C. Yin., M. El-Bishouty, and Y. Yano, "Computer supported ubiquitous learning environment for vocabulary learning," *International Journal of Learning Technology*, vol. 5, no. 1, pp. 5-24, 2010.
- [26] İ. Yaman and E. Ekmekçi, "A Shift from CALL to MALL?," *Participatory Educational Research (PER)*, vol. 4, no. Special Issue, pp. 25-32, 2016.
- [27] A. Kukulska-Hulme and L. Shield, "An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction," *ReCALL*, vol. 20, no. 3, pp. 271-289, 2008.
- [28] R. F. Hudson, P. C. Pullen, H. Lane, L. Joseph, and K. Torgesen, "The Complex Nature of Reading Fluency: A Multidimensional View," *Reading and Writing Quarterly*, vol. 25, no. 1, pp. 4-32, 2009.
- [29] M. Yıldız, K. Yıldırım, S. Ateş, and T. Rasinsky, "Perceptions of Turkish parents with children identified as dyslexic about the problems that they and their children experience," *Reading Psychology*, vol. 33, no. 5, pp. 399-422, 2012.
- [30] Y.-S. G. Kim, "Developmental, Component-Based Model of Reading Fluency: An Investigation of Predictors of Word- Reading Fluency, Text- Reading Fluency, and Reading Comprehension," *Reading Research Quarterly*, vol. 50, no. 4, pp. 459-481, 2015.
- [31] L. S. Fuchs, D. Fuchs, M. K. Hosp, and J. R. Jenkins, "Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis," *Scientific Studies of Reading*, vol. 5, no. 3, pp. 239-256, 2001.
- [32] S. J. Samuels, "Toward a Model of Reading Fluency," in *What research has to say about fluency instruction*, S. J. Samuels and A. E. Farstrup, Eds.: International Reading Association, 2006, pp. 24-46.
- [33] M. Wolf and T. Katzir-Cohen, "Reading fluency and its intervention," *Scientific Studies of Reading*, vol. 5, no. 3, pp. 211-239, 2001.
- [34] D. LaBerge and S. J. Samuels, "Towards a theory of automatic information processing in reading," *Cognitive Psychology*, vol. 6, pp. 293-323, 1974.
- [35] T. V. Rasinski et al., "Impact of Classroom-Based Fluency Instruction on Grade One Students in an Urban Elementary School," *Education Sciences*, vol. 10, pp. 227-236, 2020.
- [36] B. Alharbi, "The Impact of Glossed Texts on Reading Comprehension among Tertiary Saudi Students," *English Language Teaching*, vol. 11, no. 3, pp. 153-161, 2018.
- [37] M. Gutiérrez-Colón, A. D. Frumuselu, and H. Curell, "Mobile-assisted Language learning to enhance L2 reading comprehension: A selection of implementation studies between 2012-2017," *Interactive Learning Environments*, pp. 1-9, 2020.
- [38] H. Alshenqeeti and M. Alrahaili, "Effects of Task Repetition on Saudi EFL Learners' Reading Comprehension," *Arab World English Journal (AWEJ) Volume*, vol. 11, 2020.
- [39] M. Khojah and M. Thomas, "Smartphone-mediated EFL Reading Tasks: A Study of Female Learners' Motivation and Behaviour in Three Saudi Arabian Classrooms," *Asian EFL Journal*, vol. 25, no. 2, 2021.
- [40] I. Khan, A. H. Ibrahim, A. Kassim, and R. M. I. Khan, "Evaluating The Efficacy Of Active Reading Software In Enhancing EFL Learners' Reading Comprehension Skills," *International Journal of Scientific & technology Research* vol. 8, no. 12, pp. 1861-1869, 2019.
- [41] E. Bensalem, "Foreign Language Reading Anxiety in the Saudi Tertiary EFL Context," 2020.
- [42] I. Hassan Taj, F. Ali, M. Sipra, and W. Ahmad, "Effect of technology enhanced language learning on vocabulary acquisition of EFL learners," *International Journal of Applied Linguistics & English Literature*, vol. 6, no. 3, 2017.
- [43] A. A. Abanomey, "Do EFL Saudi learners perform differently with online reading? An exploratory study," *Journal of King Saud University - Languages and Translation*, vol. 25, no. 1, pp. 1-11, 2013/01/01/ 2013.
- [44] R. Alshammari, M. Parkes, and R. Adlington, "Using WhatsApp in EFL instruction with Saudi Arabian university students," *Arab World English Journal (AWEJ) Volume*, vol. 8, 2017.
- [45] W. Albiladi and K. Alshareef, "The Use of Tablets in English Teaching in Saudi Arabia: Implications and Challenges," *International Journal of Language and Education*, vol. 7, no. 3, pp. 280-294, 07/01 2018.
- [46] H. Alotaibi, "Decide, Design, Develop and Evaluate: Computer-Based reading lessons for ESL learners " presented at the ICT for Language Learning, Florence, Italy, 2012.
- [47] B. Varol and G. Erçetin, "Effects of Working Memory and Gloss Type on L2 Text Comprehension and Incidental Vocabulary Learning in Computer-Based Reading," *Procedia - Social and Behavioral Sciences*, vol. 232, pp. 759-768, 2016/10/14/ 2016.
- [48] M. A. Ali, "Reading in a foreign language effectiveness of computer-based reading instruction in comparison to teacher-based reading instruction," PhD Doctoral dissertation., Loughborough University, 2004.
- [49] E. Mahmoudi, A. b. A. Samad, and N. Z. B. A. Razak, "Attitude and Students' Performance in Computer Assisted English Language Learning (CAELL) for Learning Vocabulary," *Procedia - Social and Behavioral Sciences*, vol. 66, pp. 489-498, 2012/12/07/ 2012.
- [50] Y.-J. Lan, Y.-T. Sung, and K.-E. Chang, "A Mobile-Device-Supported Peer-Assisted Learning System for Collaborative Early EFL Reading," *Language Learning & Technology*, vol. 11, no. 3, pp. 130-151, 2007.
- [51] C.-M. Chen, M.-C. Li, and T.-C. Chen, "A web-based collaborative reading annotation system with gamification mechanisms to improve reading performance," *Computers & Education*, vol. 144, p. 103697, 2020/01/01/ 2020.
- [52] S.-J. Yang, C.-P. Lin, Mei-HwaYang, Yin-JuanShao, and WenliChen, "Computer-supported Collaborative Learning for Elementary School Students on the Effectiveness of Reading Comprehension," in *Proceedings of the 19th International Conference on Computers in Education Chiang Mai, Thailand, 2011: Asia-Pacific Society for Computers in Education*.
- [53] K. Ae-Hwa et al., "Improving the Reading Comprehension of Middle School Students With Disabilities Through Computer-Assisted Collaborative Strategic Reading," *Remedial and Special Education - REM SPEC EDUC*, vol. 27, 07/01 2006.

- [54] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied Thematic Analysis*. London: SAGE Publications, 2012.
- [55] M. Crowe, M. Inder, and R. Porter, "Conducting qualitative research in mental health: Thematic and content analyses," *Australian & New Zealand Journal of Psychiatry*, vol. 49, no. 7, pp. 616-623, 2015.
- [56] R. E. Boyatzis, *Transforming Qualitative Information: Thematic Analysis and Code Development*. SAGE Publications, 1998.
- [57] M. Schreier, *Qualitative Content Analysis in Practice*. London: SAGE, 2012.
- [58] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. London: SAGE, 2004.
- [59] M. Amanfi, *The 4-Step Thematic Data Analysis With MAXQDA*. Amazon Digital Services LLC - KDP Print US, 2019.
- [60] S. Friese, *Qualitative Data Analysis with ATLAS.ti*. London: SAGE, 2014.
- [61] N. H. Woolf and C. Silver, *Qualitative Analysis Using ATLAS.ti: The Five-Level QDATM Method*. London; New York: Routledge, 2017.
- [62] N. H. Woolf and C. Silver, *Qualitative Analysis Using ATLAS. ti: The Five-Level QDATM Method*. Routledge, 2017.
- [63] C. H. D. Larenas, P. A. Hernandez, and M. O. Navarrete, "A case study on EFL teachers' beliefs about the teaching and learning of English in public education," *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*, no. 23, pp. 171-186, 2015.
- [64] T. Gupta, "A Study of Problem Solving Strategies Using ATLAS. ti," in *Computer-Aided Data Analysis in Chemical Education Research (CADACER): Advances and Avenues: ACS Publications*, 2017, pp. 133-155.
- [65] J. d. L. Ferreira and G. R. Bertotti, "Continuing Education for Professional Development in Higher Education Teaching," *Creative Education*, vol. 7, pp. 1425-1435, 2016.
- [66] M. Alshahrani, "A Brief Historical Perspective of English in Saudi Arabia," *Journal of Literature, Languages and Linguistics*, vol. 26, pp. 43-47, 2016.
- [67] S. Faruk, "English language teaching in Saudi Arabia: A world system perspective," *Scientific Bulletin of the Politehnica University of Timișoara Transactions on Modern Languages*, vol. 12, no. 1-2, pp. 73-80, 2013.
- [68] P. Hubbard, "Educating the CALL specialist," *International Journal of Innovation in Language Learning and Teaching*, vol. 3, no. 1, pp. 3-15, 2009.
- [69] N. Presson, C. Davy, and B. MacWhinney, "Experimentalized CALL for adult second language learners," J. W. Schwieter, Ed. msterdam; Philadelphia: John Benjamins Publishing Company, 2013, pp. 139-164.
- [70] T. Heift and N. Vyatkina, "Technologies for Teaching and Learning L2 Grammar," in *The Handbook of Technology and Second Language Teaching and Learning*, 2017, pp. 26-44.

Optimal Allocation of DG and D-STATCOM in a Distribution System using Evolutionary based Bat Algorithm

Surender Reddy Salkuti

Department of Railroad and Electrical Engineering
Woosong University, Daejeon, South Korea

Abstract—In this work, a methodology to find the optimal allocation (i.e., sizing and location) of Distributed Generators (DGs) and Distribution-static compensators (D-STATCOM) in a radial distribution system (RDS) is proposed. Here, the voltage stability index (VSI) is utilized to find the optimal location for the D-STATCOM, and loss sensitivity factor (LSF) method is utilized to find the optimal location for distributed generation. In this work, the proposed work is formulated as a non-linear optimization problem and it is solved using the meta-heuristic/evolutionary-based algorithm. The evolutionary-based Bat algorithm is used to find optimal sizes of D-STATCOM and DGs in RDSs. To check the validity and feasibility and validity of the proposed optimal allocation approach, two standard IEEE 34 and 85 bus RDSs are considered in this paper. The simulation results show reduction in power losses and enhancement in bus voltages in the RDSs.

Keywords—Bat algorithm; distributed generation; voltage stability index; loss sensitivity; optimal location and size; radial distribution system

NOMENCLATURE

V_i	Voltage magnitude at i^{th} bus.
BIBC	Bus current Injection to Branch Current matrix.
$iter$	Iteration number.
$iter_{max}$	Maximum number of iterations.
N_b	Number of buses.
V_{i+1}	Voltage magnitude at $(i+1)^{\text{th}}$ bus.
ϵ	Equally distributed number between -1 and 1.
A_i^k	Average loudness of all bats in the k^{th} iteration.
Z_i	Pulse emission rate of i^{th} Bat.
ω	Inertia weight factor.
$R_{i,(i+1)}$	Resistance between the buses i and $(i+1)$.
$X_{i,(i+1)}$	Reactance between the buses i and $(i+1)$.
P_{Di}, Q_{Di}	Active and reactive power demands at i^{th} bus.
I_L	Current flow in the line.
P_{DS}	Power supplied from the RDS.
$P_{D-STATCOM}$	Active power from the D-STATCOM.
S^*	Global solution.
P_{DG}	Power output from the DG.
r	Pulse emission rate.
β	Random number between 0 and 1.
f^{min}, f^{max}	Minimum and maximum frequencies.
V_i^{min}, V_i^{max}	Lower and upper voltages at i^{th} bus.

I. INTRODUCTION

In recent years, renewable energy is gaining huge attention throughout the world because of several reasons. The increased awareness towards renewable and sustainable energies, such as wind and solar energies has stimulated the trend towards energy-efficient devices and energy conservation techniques to optimize the energy demand and cost-saving. The distribution networks acquire higher losses (i.e., around 13%) in the entire power system, and the voltage stability issues of radial distribution systems (RDSs) have to be considered urgently. From the planning perspective, the optimal operating of the distribution system is used for shunt volt-ampere reactive planning, series capacitor planning, transfer capability Studies, reactive interchange studies, loss optimization studies. Distributed generation (DG) from various sources offers an economic and reliable source of electricity to the customers. Optimal integration of distributed generation into an existing power network plays a vital role because of its benefits. Optimal DG allocation (i.e., location and sizing) is the challenging issue to improve system efficiency by reducing the power losses and by improving the voltage stability and security of the system. D-STATCOM is used for reactive power compensation, which in turn minimizes voltage drop and power losses that are caused by the increasing demands [1]. The optimal solution is obtained after solving several power flow solutions, corresponding to a specified set of consumer demands.

With the advancement of smart grid technologies, managing the distribution network safely and economically has been one of the major business activities of an electrical utility. One of the fundamental issues is that the energy transferred by the network is continually changing as the consumer load changes; in addition to this, the network itself is not static [2]. Long-term planning studies are carried out for several years ahead to check the resilience of the existing networks for various situations. Power system analysis tools including optimization studies are used to design construction or refurbishment of plant for further expansion. Although an optimal running arrangement for an expected load pattern may be determined in a planning study, this may not match the system conditions on that day [3].

The author in [4] illustrates the importance of optimal sizing and location of DG in RDS. The author in [5] utilizes

an analytical-based method for finding the optimal size and location of D-STATCOM and DG placement in RDS. A methodology for the optimal sizing D-STATCOM in RDSs considering load demand uncertainty is presented in [6]. A fuzzy-based approach for feeder reconfiguration considering D-STATCOM to reduce operating cost and active power losses in RDSs has been proposed in [7]. An optimal placement approach using multi-objective-based sensitivity methodology and evolutionary-based technique for determining the optimal allocation of D-STATCOM in unbalanced RDS is described in [8]. An approach to find the optimal sizing of DGs in RDSs with uncertain topologies is solved using Monte Carlo Simulation (MCS) is proposed in [9]. An approach to finding optimal sizing and placement of DG and D-STATCOM is analyzed in [10] to optimize power loss, costs, and voltage profile enhancement in the RDS.

Nowadays, evolutionary techniques have been used for different applications of distribution systems (DSs). An approach to find the optimal location and sizing of D-STATCOM and DG to optimize power losses of the system using Harmony search algorithm (HSA) is presented in [11], using Cuckoo search algorithm (CSA) is presented in [12], using invasive weed optimization technique is presented in [13], using Stud Krill herd technique is presented in [14], using backtracking search optimization (BSO) technique is presented in [15], using Ant Lion optimization (ALO) algorithm is presented in [16], using improved differential search algorithm is presented in [17], using whale optimization algorithm (WOA) is described in [18]-[19], using the particle swarm optimization (PSO) is presented in [20].

An efficient optimization approach is required for the adjustment of control variables in the optimization. Due to various limitations of classical optimization algorithms, in the present work evolutionary algorithms are used as an optimization tool. Artificial neural networks find applications in power system planning, operation, and analysis. Some of the applications are in load forecasting, optimal power flow, unit commitment, state estimation, static and dynamic security assessment, fault detection, fault location, system voltage stability assessment. Fuzzy systems find applications in power system planning, operation, and control. A few of the applications include generation expansion planning, reliability analysis, daily load forecasting, voltage, and reactive power control, state estimation, security assessment, fault diagnosis, converters control, and various control systems design. In this work, bat algorithm (BA) is used for solving the optimal location and sizing of D-STATCOMs and DGs in RDSs. The voltage stability index (VSI) is considered for finding the optimal location to install D-STATCOM, and loss sensitivity factor (LSF) method for determining an optimal location to install DG.

II. PROBLEM FORMULATION

A. Distribution Load Flow (DLF)

Here, a direct approach is utilized to solve the distribution load flow (DLF), which gives the voltage at each bus and total power losses in the RDS. Fig. 1 depicts the single-line diagram (SLD) of two buses in RDS [21].

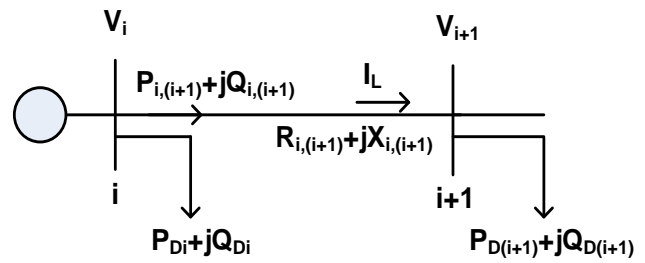


Fig. 1. SLD of Two Buses in Radial Distribution System (RDS).

From Fig. 1, voltage at the bus (i+1) is expressed as [22],

$$V_{i+1} = V_i - I_L [R_{i,(i+1)} + jX_{i,(i+1)}] \quad (1)$$

The branch current I_L is expressed as

$$I_L = [BIBC][i] \quad (2)$$

From Fig. 1, the injected current at the i^{th} bus is expressed as [23],

$$I_L = \left(\frac{P_{Di} + jQ_{Di}}{V_i} \right)^* \quad (3)$$

The active power loss between buses i and $(i+1)$ can be calculated by using the load flow solution and it is expressed as [24],

$$P_{loss(i,(i+1))} = \left(\frac{P_{i,(i+1)}^2 + Q_{i,(i+1)}^2}{|V_{i,(i+1)}|^2} \right) R_{i,(i+1)} \quad (4)$$

Total power loss (P_{TL}) of RDSs can be found by adding losses in all the lines, and it can be expressed as [25],

$$P_{TL} = \sum_{i=1}^{N_b} P_{loss(i,(i+1))} \quad (5)$$

The D-STATCOM also injects the reactive power into the power network. The detailed modeling of D-STATCOM incorporated in the load flow study has been presented in [26]. The detailed description of VSI is presented in references [27-28], and the description of LSF is presented in references [29-30].

B. Objective Function: Minimization of Active Power Losses

The main concentration of the work is the development of more robust and faster optimal allocation approach. Load flow solution is a part of optimal power flow solution which is considered as the equality constraint of the optimal allocation problem. Power flow is the most fundamental numerical algorithm for the analysis of distribution network. The cost minimization objective is mainly influenced by the variables that mostly influence the cost minimization are active power generations of generators, voltage magnitude of generators. For loss minimization objective the active power generations almost at maximum values. Voltage magnitudes of generators, transformer taps and shunts all influence loss. This objective function plays a major role in RDSs. This objective function can be formulated as [31],

$$\text{minimize } J_1 = \text{minimize } (P_{TL}) \quad (6)$$

C. Equality Constraints

This constraint represents the power balance among the total power generation from the DS, DG, and D-STATCOM, and the system load plus losses. Mathematically, this constraint can be modeled as [32],

$$\sum_{i=1}^{N_b} P_{Di} + P_{TL} = P_{DG} + P_{D-STATCOM} + P_{DS} \quad (7)$$

D. Inequality Constraints

These constraints on control parameters are so selected that they fall within the permissible limits. Then the functional inequality constraints are considered using the penalty method, where the objective function is augmented by a penalty term, for each violated functional constraint. The limitation of the method is in the modeling of transformer taps [33].

1) Constraint on Bus Voltage Magnitudes

Voltage magnitude at each bus is restricted by [34],

$$V_i^{min} \leq |V_i| \leq V_i^{max} \quad (8)$$

2) Reactive Power Compensation (RPC) Constraint

Reactive power ($Q_{D-STATCOM}$) injected by the D-STATCOM at i^{th} candidate bus is limited by [35],

$$Q_{D-STATCOM,i}^{min} \leq Q_{D-STATCOM,i} \leq Q_{D-STATCOM,i}^{max} \quad (9)$$

3) Active Power Constraint

Real power injected by DG at i^{th} bus is limited by [36, 37],

$$P_{DG,i}^{min} \leq P_{DG,i} \leq P_{DG,i}^{max} \quad (10)$$

III. BAT ALGORITHM

The BA is inspired by echolocation ability of the microbats [38]. Suppose a Bat has a randomly generated position of S_i and the velocity of V_i . Here, the Bat is selected based on the sound pulses that emits with a loudness of A_0 , charging wavelength of λ , the fixed minimum frequency of f_{min} , and r is a random number that takes a value in the range [0, 1].

BA develops the potential candidate solutions as the virtual microbats with velocities (V_i), positions (S_i), and frequencies (f_i), and they are restructured as [39],

$$f_i = (f_{max} - f_{min})\beta + f_{min} \quad (11)$$

$$V_i^{k+1} = V_i^k + (S_i^k - S^*)f_i \quad (12)$$

where k is number of iterations, and i is number of Bats.

$$S_i^{k+1} = S_i^k + V_i^{k+1} \quad (13)$$

The new position is updated using,

$$S_{i,new}^k = S_{i,old}^k + \varepsilon A_i^k \quad (14)$$

In each iteration/generation, A_i and Z_i are formulated as,

$$A_i^{k+1} = \rho A_i^k \quad (15)$$

$$Z_i^{k+1} = Z_i^0 (1 - e^{-\delta k}) \quad (16)$$

where δ and ρ are the constants, where $0 < \rho < 1$, $\delta > 0$. Bat's velocity is updated using ω , and it can be represented by [40],

$$V_i^{k+1} = \omega V_i^k + (S^* - S_i^k)f_i \quad (17)$$

The flow chart of the Bat algorithm (BA) for the implementation of proposed optimization problem is depicted in Fig. 2.

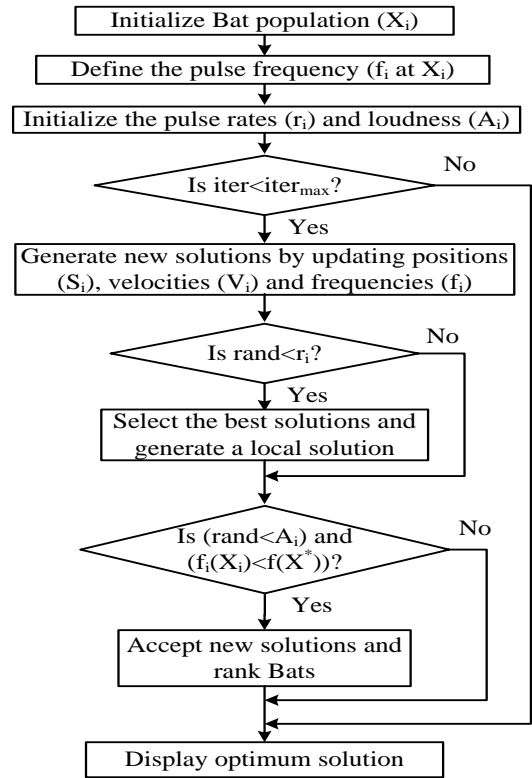


Fig. 2. Flow Chart of Bat Algorithm (BA).

Implementation of the proposed methodology is described below:

- Step 1: Input the bus data, load data of RDS, data related to Bat algorithm.
- Step 2: Perform the distribution load flow (DLF) for the base case condition (i.e., without DG and D-STATCOM), and find bus voltages, total power losses, loss sensitivity factors (LSF), and voltage stability index (VSI).
- Step 3: Find optimal locations for placing DGs using the LSF approach and D-STATCOM using VSI.
- Step 4: Initialize the random Bat algorithm (X_i) for the amount of kW's and kVAr's to be injected within their minimum and maximum limits.
- Step 5: Generate a new solution by adjusting the frequencies (f_i) using equation (11). Update positions (S_i^{k+1}) and velocities (v_i^{k+1}) using the equations (12) and (13).

- Step 6: Calculate fitness function using objective function (i.e., Equation (6)) for each Bat location.
- Step 7: Generate a random number using the *rand* function. If $\text{rand} > \text{pulse rates } (r_i)$, then determine a local solution near the best solution using the equation (17). Otherwise, go to Step 10.
- Step 8: Check if $(\text{rand} < A_i)$ and $f(X_i) < f(X^*)$. If YES, then accept new solutions, reduce A_i , and increase r_i then update the current best solution. If NO, then set the best solution as the current best solution.
- Step 9: Check the stopping criteria. If $(\text{iter} < \text{iter}_{max})$, then go to Step 5, else STOP the algorithm.
- Step 10: Print the optimal solution.

IV. SIMULATION RESULTS AND DISCUSSION

In this work, IEEE 34 bus and 85 bus RDSs are selected to check the validity of proposed optimal allocation methodology. For each RDS, 3 case studies are simulated, and they are:

- Case 1: Base case condition, i.e., system without DG and D-STATCOM.
- Case 2: System with single DG and D-STATCOM.
- Case 3: System with multiple DGs and D-STATCOMs.

A. Results on IEEE 34 bus RDS

The bus data and line data of 34 buses, 33 branch RDS is selected from [41].

1) *Case study 1: Base Case (without D-STATCOM and DG):* In this case study, the base case is considered without installing the D-STATCOM and DG. Here, the obtained optimal real power loss is 221.67 kW and reactive power loss is 65.1 kVAr, and they are presented in Table I. The voltage profile of 34 bus RDS for case study 1 is depicted in Fig. 3. The minimum voltage in the system is 0.9355 p.u. and it is at bus number 27. And also minimum VSI obtained is 0.7875 p.u.

2) *Case study 2: Network with single DG and D-STATCOM:* Here, single D-STATCOM and DG are installed simultaneously at optimal places to achieve optimum objectives. Simulation results obtained in this case are presented in Table I. Here, the obtained optimal placement of DG at bus number 21 and its optimal size is 2154.75 kW. By using the proposed approach, bus number 22 is obtained as the suitable location for installing the D-STATCOM, and its size is 1250.8 kVAr. Here, the obtained optimal active power loss is 56.45 kW and it is 74.53% less compared to the loss obtained in case study 1. Table I and Fig. 3 depict the optimal allocation of DG and D-STATCOMs, and voltage profile at each bus. In this case, the minimum voltage obtained is 0.9756 p.u., which has been improved, compared to case study 1 (i.e., 0.9355 p.u.). In similar lines, minimum VSI obtained in this case is 0.9118 p.u. which is also improved as compared to case study 1 (i.e., 0.7875 p.u.).

TABLE I. OBTAINED RESULTS FOR IEEE 34 BUS RDS

	Case study 1	Case study 2	Case study 3
Optimal size and location of DGs	---	2154.75 kW at bus 21	1820.53 kW at bus 21 763.11 kW at bus 32 102.07 kW at bus 34
Optimal size and location of D-STATCOM	---	1250.8 kVAr at bus 22	960.82 kVAr at bus 18 618.51 kVAr at bus 22
V_{min} (p.u.)	0.9355	0.9756	0.9912
VSI_{min} (p.u.)	0.7875	0.9118	0.9701
Optimum power loss (kW)	221.67	56.45	18.94
Loss reduction (%)	---	74.53	91.46

3) *Case study 3: System with multiple DGs and D-STATCOMs:* Here, multiple D-STATCOMs and DGs are allocated in the system and results are reported in Table I. In this case study, it has been considered that three DGs and two D-STATCOMs are placed in the RDS. The suitable locations of DGs obtained are at buses 21, 32, 34, and their optimal sizes are 1820.53 kW, 763.11 kW, and 102.07 kW, respectively. Optimal locations obtained for D-STATCOM are buses 18, 22 and their optimal sizes are 960.82 kVAr, 618.51 kVAr, respectively. The obtained minimum power loss is 18.94 kW and it is 91.46% less when compared to case study 1 and 66.45% less when compared to case study 2. Minimum voltage obtained is 0.9912 p.u., which has been improved from case study 1 (i.e., 0.9355 p.u.) and case study 2 (i.e., 0.9756 p.u.). Minimum VSI obtained is 0.9701 p.u., which has been improved from case study 1 (i.e., 0.7875 p.u.) and case study 2 (i.e., 0.9118 p.u.).

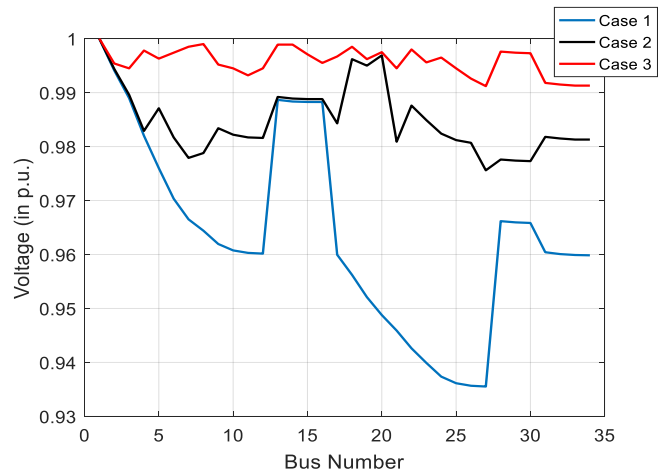


Fig. 3. Voltage Profile of 34 Bus RDS for Cases 1, 2, and 3.

B. Results on IEEE 85 bus RDS

The bus data and line data of 85 bus, 84 branch RDS is selected from [42].

1) *Case study 1: Base Case (without D-STATCOM and DG):* In this case study, the base case is considered without installing the D-STATCOM and DG. The obtained optimal

active power loss is 315.70 kW and reactive loss is 198.356 kVAr, and they are presented in Table II. Voltage profile of 85 bus RDS for case study 1 is depicted in Fig. 4. The minimum voltage in the system is 0.8714 p.u. and it is a bus number 54. And also the obtained minimum VSI is 0.6713 p.u.

TABLE II. OBTAINED RESULTS FOR IEEE 85 BUS RDS

	Case 1	Case 2	Case 3
Optimal size and location of DG	---	2371.4 kW at bus 55	359.4 kW at bus 8 680.3 kW at bus 29 2250.7 at bus 55
Optimal size and location of D-STATCOM	---	1185.6 kVAr at bus 58	692.1 kVAr at bus 27 1079.8 kVAr at bus 58
V_{min} (p.u.)	0.8714	0.9346	0.9704
VSI_{min} (p.u.)	0.6713	0.8631	0.9240
Optimum power loss (kW)	315.70	82.36	31.94
Loss reduction (%)	---	73.91	89.88

2) *Case study 2: System with single D-STATCOM and DG:* Here, single DG and D-STATCOMs are located simultaneously at optimal places to achieve optimum objectives, i.e., improving the voltage profile or minimizing the power losses. The obtained results are presented in Table II. Here, the obtained optimal placement of DG at bus number 55 and its optimal size is 2371.4 kW. By using the proposed approach, bus number 58 is obtained as optimal location for D-STATCOM, and its optimal size is 1185.6 kVAr. The optimal active power loss obtained is 82.36 kW and it is 82.36% less compared to the loss obtained in case study 1. Table II and Fig. 4 depict optimal sizes and locations of DG and D-STATCOMs and the voltage profile at each bus. The obtained minimum voltage is 0.9346 p.u., which has been improved, compared to case study 1 (i.e., 0.8714 p.u.). In similar lines, the obtained minimum VSI is 0.8631 p.u. which is also improved as compared to case study 1 (i.e., 0.6713 p.u.).

3) *Case study 3: System with multiple DGs and D-STATCOMs:* Here, multiple D-STATCOMs and DGs are located in the system and the results are reported in Table II. In this case study, it has been considered that three DGs and two D-STATCOMs are placed in the network. Optimal locations of DGs obtained are at buses 8, 29, 55, and their optimal sizes are 359.4 kW, 680.3 kW, and 2250.7 kW, respectively. Optimal locations of D-STATCOM are at buses 27, 58 and their optimal sizes are 692.1 kVAr, 1079.8 kVAr, respectively. Here, the obtained minimum loss is 31.94 kW and it is 89.88 % less when compared to case study 1 and 61.22% less when compared to case study 2. Minimum voltage obtained is 0.9704 p.u., and it has been improved from case study 1 (i.e., 0.8714 p.u.) and case study 2 (i.e., 0.9346 p.u.). Minimum VSI obtained is 0.9240 p.u., which has been improved from case study 1 (i.e., 0.6713 p.u.) and case study 2 (i.e., 0.8631 p.u.).

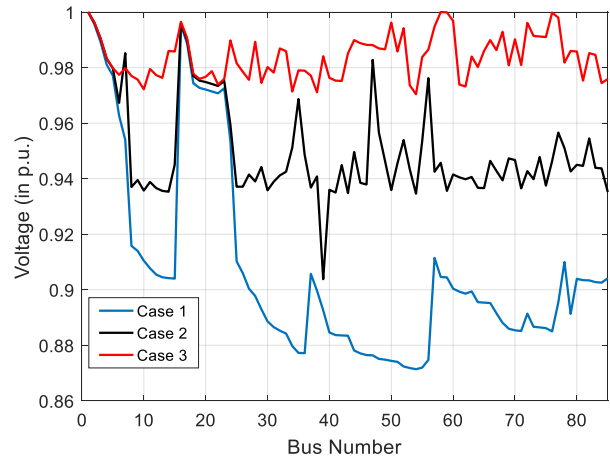


Fig. 4. Voltage Profile of 85 Bus RDS for Cases 1, 2, and 3.

V. CONCLUSIONS

Optimal allocation that is sizing and location of D-STATCOM and DG problem have been solved in this paper to optimize system losses and to enhance voltage profile of the radial distribution systems (RDSs). In this work, loss sensitivity factor (LSF) method is utilized to find suitable location for the DG, and voltage stability index (VSI) is utilized to determine the suitable location for the D-STATCOM. Bat algorithm (BA) is used for determining optimal sizing of DG and D-STATCOM in RDSs. The validity of the proposed optimization problem is examined on IEEE 34 bus and 85 bus RDSs. Simultaneous allocation of DGs and D-STATCOMs in RDSs reduces power losses and enhances the voltage profile in the RDS. Future work can be extended for finding the optimal locations for the electric vehicles charging stations.

ACKNOWLEDGMENT

This research work was funded by “Woosong University’s Academic Research Funding – 2021”.

REFERENCES

- [1] Y. Thangaraj, R. Kuppan, “Multi-objective simultaneous placement of DG and DSTATCOM using novel lightning search algorithm”, *Journal of Applied Research and Technology*, vol. 15, pp. 477-491, 2017.
- [2] T.P. Nguyen, D.N. Vo, T.T. Tran, “Optimal Number, Location, and Size of Distributed Generators in Distribution Systems by Symbiotic Organism Search Based Method”, *Advances in Electrical and Electronic Engineering*, vol. 15, no. 5, pp. 724-735, Dec. 2017.
- [3] S.R. Salkuti, “Optimal Location and Sizing of Shunt Capacitors with Distributed Generation in Distribution Systems”, *ECTI Transactions on Electrical Engineering, Electronics, and Communications*, vol. 19, no. 1, pp. 34-42, Feb. 2021.
- [4] S.S. Kola, “A Review on Optimal Allocation and Sizing Techniques for DG in Distribution Systems”, *International Journal of Renewable Energy Research*, vol. 8, no. 3, pp. 1236-1256, Sept. 2018.
- [5] B. Weqar, M.T. Khan, A.S. Siddiqui, “Optimal Placement of Distributed Generation and D-STATCOM in Radial Distribution Network”, *Smart Science*, vol. 6, no. 2, pp. 125-133, 2018.
- [6] E. Shahryari, H. Shayeghi, M.M. oradzadeh, “Probabilistic and Multi-Objective Placement of D-STATCOM in Distribution Systems Considering Load Uncertainty”, *Electric Power Components and Systems*, vol. 46, no. 1, pp. 27-42, 2018.
- [7] M. Mohammadi, M. Abasi, A.M. Rozbahani, “Fuzzy-GA based algorithm for optimal placement and sizing of distribution static

- compensator (DSTATCOM) for loss reduction of distribution network considering reconfiguration”, *Journal of Central South University*, vol. 24, no. 2, pp. 245-258, Feb. 2017.
- [8] X. Su, H. Liu, Y. Fu, P. Wolfs, “Multi-objective DSTATCOM placement based on sensitivity analysis and genetic algorithm in unbalanced mv distribution networks”, *IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*, Auckland, 2017, pp. 1-5.
- [9] C.B. Donadel, J.F. Fardin, L.F. Encaracao, “Optimal Placement of Distributed Generation Units in a Distribution System with Uncertain Topologies using Monte Carlo Simulation”, *International Journal of Emerging Electric Power Systems*, vol. 16, no. 5, pp. 431-442, Oct. 2015.
- [10] K.R. Devabalaji, K. Ravi, “Optimal size and siting of multiple DG and DSTATCOM in radial distribution system using Bacterial Foraging Optimization Algorithm”, *Ain Shams Engineering Journal*, vol. 7, pp. 959-971, 2016.
- [11] T. Yuvaraj, K.R. Devabalaji, K. Ravi, “Optimal placement and sizing of DSTATCOM using Harmony Search algorithm”, *Energy Procedia*, vol. 79, pp. 759-765, 2015.
- [12] T. Yuvaraj, K. Ravi, K.R. Devabalaji, “Optimal allocation of DG and DSTATCOM in radial distribution system using Cuckoo Search optimization algorithm”, *Modelling and Simulation in Engineering*, vol. 2017, pp. 1-11, 2017.
- [13] D.R. Prabha, T. Jayabarathi, “Optimal placement and sizing of multiple distributed generating units in distribution networks by invasive weed optimization algorithm”, *Ain Shams Engineering Journal*, vol. 7, pp. 683-694, 2016.
- [14] S.A.C. Devi, L. Lakshminarasimman, R. Balamurugan, “Stud Krill herd Algorithm for multiple DG placement and sizing in a radial distribution system”, *Engineering Science and Technology, an International Journal*, vol. 20, pp. 748-759, 2017.
- [15] A.E. Fergany, “Optimal allocation of multi-type distributed generators using backtracking search optimization algorithm”, *International Journal of Electrical Power Energy Systems*, vol. 64, pp. 1197-1205, 2015.
- [16] P.D.P. Reddy, V.C.V. Reddy, T.G. Manohar, “Ant Lion optimization algorithm for optimal sizing of renewable energy resources for loss reduction in distribution systems”, *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 663-680, Dec. 2018.
- [17] S.K. Injeti, “A Pareto optimal approach for allocation of distributed generators in radial distribution systems using improved differential search algorithm”, *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 908-927, Dec. 2018.
- [18] P.D.P. Reddy, V.C.V. Reddy, T.G. Manohar, “Optimal renewable resources placement in distribution networks by combined power loss index and whale optimization algorithms”, *Journal of Electrical Systems and Information Technology*, vol. 5, no. 2, pp. 175-191, Sept. 2018.
- [19] D.B. Prakash, C. Lakshminarayana, “Optimal siting of capacitors in radial distribution network using Whale Optimization Algorithm”, *Alexandria Engineering Journal*, vol. 56, no. 4, pp. 499-509, Dec. 2017.
- [20] D.B. Prakash, C. Lakshminarayana, “Multiple DG Placements in Distribution System for Power Loss Reduction Using PSO Algorithm”, *Procedia Technology*, vol. 25, pp. 785-792, 2016.
- [21] F. Iqbal, M.T. Khan, A.S. Siddiqui, “Optimal placement of DG and DSTATCOM for loss reduction and voltage profile improvement”, *Alexandria Engineering Journal*, vol. 57, no. 2, pp. 755-765, Jun. 2018.
- [22] S.R. Salkuti, “Optimal location and sizing of DG and D-STATCOM in distribution networks”, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 3, pp. 1107-1114, Dec. 2019.
- [23] X. Su, H. Liu, Y. Fu, P. Wolfs, “Multi-objective DSTATCOM placement based on sensitivity analysis and genetic algorithm in unbalanced mv distribution networks”, *IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*, Auckland, New Zealand, 2017, pp. 1-5.
- [24] N. Mohandas, R. Balamurugan, L. Lakshminarasimman, “Optimal location and sizing of real power DG units to improve the voltage stability in the distribution system using ABC algorithm united with chaos”, *Electric Power Energy Systems*, vol. 66, pp. 41-52, 2015.
- [25] P.D.P. Reddy, V.C.V. Reddy, T.G. Manohar, “Application of flower pollination algorithm for optimal placement and sizing of distributed generation in Distribution systems”, *Journal of Electrical Systems and Information Technology*, vol. 3, no. 1, pp. 14-22, May 2016.
- [26] M. Junjie, W. Yulong, L. Yang, “Size and Location of Distributed Generation in Distribution System Based on Immune Algorithm”, *Systems Engineering Procedia*, vol. 4, pp. 124-132, 2012.
- [27] S. Sudabattula, M. Kowsalya, “Optimal allocation of wind based distributed generators in distribution system using Cuckoo Search Algorithm”, *Procedia Computer Science*, vol. 92, pp. 298-304, 2016.
- [28] M.J.H. Moghaddam, S. Arabi-Nowdeh, M. Bigdeli, D. Azizian, “A multi-objective optimal sizing and siting of distributed generation using ant lion optimization technique”, *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2101-2109, Dec. 2018.
- [29] S. Kansal, V. Kumar, B. Tyagi, “Hybrid approach for optimal placement of multiple DGs of multiple types in distribution networks”, *International Journal of Electrical Power and Energy Systems*, vol. 75, pp. 226-235, 2016.
- [30] B. Das, V. Mukherjee, D. Das, “DG placement in radial distribution network by symbiotic organism search algorithm for real power loss minimization”, *Applied Soft Computing*, vol. 49, pp. 920-936, Dec. 2016.
- [31] U. Sultana, A.B. Khairuddin, A.S. Mokhtar, N. Zareen, B. Sultana, “Grey wolf optimizer based placement and sizing of multiple distributed generation in the distribution system”, *Energy*, vol. 111, pp. 525-536, 2016.
- [32] M. Darfoun, M.E. El-Hawary, “Multi-objective optimization approach for optimal distributed generation sizing and placement”, *Electric Power Components and Systems*, vol. 43, no. 7, pp. 828-836, 2015.
- [33] N. Kaur, S.K. Jain, “Multi-Objective Optimization Approach for Placement of Multiple DGs for Voltage Sensitive Loads”, *Energies*, vol. 10, pp. 1-17, 2017.
- [34] A. Uniyal, A. Kumar, “Optimal Distributed Generation Placement with Multiple Objectives Considering Probabilistic Load”, *Procedia Computer Science*, vol. 125, pp. 382-388, 2018.
- [35] K.R. Devabalaji, K. Ravi, “Optimal size and siting of multiple DG and DSTATCOM in radial distribution system using Bacterial Foraging Optimization Algorithm”, *Ain Shams Engineering Journal*, vol. 7, no. 3, pp. 959-971, Sept. 2016.
- [36] A.Y. Abdelaziz, S.F. Mekhamer, R.H. Shehata, “Solution of distributed generation allocation problem using a novel method”, *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 2, pp. 554-564, Aug. 2019.
- [37] T. Yuvaraj, K. Ravi, “Multi-objective simultaneous DG and DSTATCOM allocation in radial distribution networks using cuckoo searching algorithm”, *Alexandria Engineering Journal*, vol. 57, no. 4, pp. 2729-2742, Dec. 2018.
- [38] H. Delkshos Abatari, M. S. S. Abad and H. Seifi, Application of bat optimization algorithm in optimal power flow, *24th Iranian Conference on Electrical Engineering (ICEE)*, Shiraz, 2016, pp. 793-798.
- [39] B.V. Rao, G.V.N. Kumar, “Optimal power flow by BAT search algorithm for generation reallocation with unified power flow controller”, *International Journal of Electrical Power & Energy Systems*, vol. 68, pp. 81-88, Jun. 2015.
- [40] S.R. Salkuti, “Bat algorithm-based back propagation approach for short-term load forecasting considering weather factors”, *Electrical Engineering*, vol. 100, pp. 1297-1303, 2018.
- [41] M. Chis, M. Salama, S. Jayaram, “Capacitor placement in distribution systems using heuristic search strategies”, *IEE Generation, Transmission and Distribution*, vol. 144, no. 3, pp. 225-230, May 1997.
- [42] M. Baran, F.F. Wu, “Optimal sizing of capacitors placed on a radial distribution system”, *IEEE Transactions on Power Delivery*, vol. 4, no. 1, pp. 735-743, Jan. 1989.

Analysis of Load Variation Consideration for Optimal Distributed Generation Placement

Aida Fazliana Abdul Kadir¹, Mohamad Fani Sulaima^{2*}, Noor Ropidah Bujal³
Mohd Nazri Bin Abd Halim⁴, Elia Erwani Hassan⁵
Faculty of Electrical Engineering, University Teknikal Malaysia Melaka
Melaka, Malaysia

Abstract—Distributed generation (DG) devices offered usefully for power losses minimization, grid reinforcement, bus voltages improvement and efficiency of a distribution system. Usually, the DG placement problem considers the predefined DG number and sizes that might result in many small DGs. However, a better solution could be reached with a minimum number of DGs, reducing installation and maintenance costs. Furthermore, the increment of load and vice versa may affect the voltage profile below or upper the tolerable limit and distribution feeders. Thus, this paper aims to analyze the impact of the variation of the load level with the DG connection in the power system by using the improved gravitational search algorithm (IGSA) as an optimization technique. The multi-objective function target reduces the total power loss, average total voltage harmonic distortion and voltage deviation in the distribution system. This study is considering six different load levels as in percentage of load. This proposed technique compares with the particle swarm optimization (PSO) and the gravitational search algorithm (GSA). This efficiency of the proposed technique tests on the 33-bus radial distribution system with six case studies.

Keywords—Distribution generation; optimization techniques; IGSA; losses minimization; optimal placement and sizing

I. INTRODUCTION

Distributed generation is a technology that generates electricity at or near to the load center by using renewable energy to produce electricity [1]. DG may provide a single structure for a residential and business, but also can be section of a microgrid. DG usually applied in industrial facilities, military base to the provided power supply, or a large university. In other terms, DG could be “electric power generation within distribution networks or on the customer side of the network” [2]. DG technologies that usually available in Malaysia is solar photovoltaic, wind power, biomass, and solar thermal systems. Nowadays, people want the energy that purifier and have less impact on the environment. They tend to pick DG as main electricity supply because DG can generate electricity with the renewable source rather than fossil fuel and, accepted through many countries due to reduction in gasses emission is primary criteria that lead for DGs implementation. Provided peak load demand, minimizes current branch loadings, voltage profile and reduces losses can be improving with better placement and sizing of DG [3]. Allocation and sizing of DG power in an inappropriate way toward the distribution network leads to power quality issues, increasing power losses, unstable power system, and rising operational cost [4]. In most research have an objective to reduce losses in

the system because the reduction in losses leads to reduction in the total cost.

Power system management has been facing major changing in the power generation sector during the past decades. Power system company must try to find the best way and solution to provide energy which is sufficient for the customer and avoid any unwanted problems in power system such as losses in the system, voltage stability and total harmonic distortion. One of the real matters found with the DG is system stability because of the interaction amongst generators and load characteristic. The increment of load or vice versa, the voltage profile drops below tolerable operating limit along with distribution feeders. Hence, the power generating station is work simultaneously, but when load increases more and more, all generating stations cannot bear the loads, and total blackout happens. Load in the distribution system affect the DGs planning significantly for the optimal placement and sizing of DG, and generally, a constant power load model is assumed in most studies [5].

Consequently, the meeting of small generation has growth and cause the rise of demand in DG utilization. The existence of DGs in the distribution system may result in some advantages such as improved power quality, voltage stability and reduction of the system, but the inappropriate installation of DGs with improper design could either cause positive and negative impact. However, it must be depending on the operational characteristic of the DGs and the criteria of the distribution network. Therefore, optimal placement and optimal of DGs is significant to be investigated for a reliable power system [6-7].

Furthermore, there is more optimization technique used in the optimal placement of DG. In [8], the researcher has investigated to reduce operation cost and decide the capacity and location of DGs in the grid by various optimization techniques. GSA has discovered better answers with fewer cost, although it used more time to stimulate the results. Moreover, the results found by GSA, in most cases, provide better results and, in all cases, [9]. In [10-12], Particle Swarm Optimization (PSO) heuristic method has been proposing to obtain the best size and best allocation for the insertion of DG within the distribution networks for active power compensation of reduction in actual power losses and enhancement in voltage profile. The whole absolute power loss reduction in the distribution system with active compensation depends on the planning of DG for maximizing the power system performance. However, in practice, the pleasant place or sizing

*Corresponding Author

will not always be possible because many constraints, i.e. due to size, may not be to be had inside the marketplace.

In another paper, the authors in [13] have presented a new optimization technique for defining optimal sizing and allocation of DG in a distribution system that was the improved gravitational search algorithm (IGSA). Its performance is compared with other heuristic methods such as PSO and GSA for optimal placement and sizing. The resulted has shown that the IGSA performs better than PSO and GSA by provided the best fitness value and the fastest average elapsed time. However, the authors do not consider the load variation environment to promote a significant losses measurement in the results reporting. In the majority of the literature, the size of DG has been decided as the DG size at the whole penetration level. However, due to the volatility of DG source like solar, the real size will be higher than the defined size at the whole penetration level [14].

Generally, the DG placement problem solves with a predefined DG number and sizes that might result in many small DGs. However, a better solution could be reached with a minimum number of DGs, cutting installation and maintenance costs. But, in the time perspective of a day, a month or a year, the active and reactive load values may experience severe changes. The operator must consider these variants. Detecting the most sensitive buses concerning the base case may not be sufficient to evaluate voltage stability margin(VSM) enhancement strongly affected by load increase that might change the system stability status. Therefore, considering the load fluctuations, a fixed size for the DG cannot guarantee the optimal power losses in the system [15-16].

Thus, in this study, different to the load variation adjustment analysis, the investigation is made in a 33-bus radial distribution system while following a similar algorithm technique. The IGSA performance compares with particle swarm optimization (PSO) and gravitational search algorithm (GSA) with six case studies. The result illustrated the losses minimization, average THDV and voltage deviation in the distribution system when load variation was considering, and the efficiency of the proposed technique in minimizing the total losses and improve the voltage deviation—the arrangement of the paper present in the following structure. Section II shows the problem formulation and limitation of the DG placement and sizing. The proposed algorithms have explained in Section III, while Section IV presents the results and discussion. Lastly, Section V concludes the finding of the study.

II. PROBLEM FORMULATION

A multi-objective optimization problem is created as a constrained non-linear integer optimization problem for solving the optimal DG placement and sizing in a distribution system. The objective is to minimize the total power loss, the average THDV and the voltage deviation. The fitness function of the optimization problem is given by.

$$F_{min} = \alpha(P_{Loss}) + \beta(V_{dev}) + \gamma(THD_V) \quad (1)$$

Where F is the fitness function, P loss is the total power loss (%), Vdev is the voltage deviation (%), and THDV is the average THDV (%) at all system busbars. At the same time, α is the coefficient factor for total power loss; β is the coefficient

factor for voltage deviation, and γ is the coefficient factor for THDV.

The total real power losses are expressed by:

$$P_{Loss} = \sum_{k=1}^n P_{Loss_k} \quad k = 1, 2, 3, 4 \dots, n \quad (2)$$

where n is the number of lines. The voltage deviation is defined by:

$$V_{dev} = \frac{V_{iref} - V_i}{V_i} \times 100\% \quad (3)$$

where V_{iref} is reference voltage at bus and V_i is the actual voltage at the bus. The average THDV is defined by:

$$THD_V = \frac{\sum_{i=1}^n THD_{Vi}}{n} \quad (4)$$

where n is the number of buses.

Generally, multi-objective methods provide a set of optimal solutions. For this paper, the sum of the coefficient factor method uses to decide the relative importance of the objectives to obtain the best optimisation solution. The coefficient factor for total power loss is assumed to be 0.4, while the average THDV and voltage deviation consider as 0.3. The coefficients are decided based on the relative importance of the objectives in order to obtain the best optimisation solution. Sum of the weights must be equal to 1 in weighted multi objective problem formulation.

A. Constraints

1) *Equality constraints*: The load flow constraints are equal to the real and reactive power flow constraints, respectively, as given below [17]:

$$P_G + PDG_i = P_{Loss} + PD_i \quad (5)$$

$$Q_G + QDG_i = Q_{Loss} + QD_i \quad (6)$$

2) *Inequality Constraints*

Power Generation Limit [11].

$$P_{DG_i}^{min} \leq P_{DG_i} \leq P_{DG_i}^{max} \quad (7)$$

$$Q_{DG_i}^{min} \leq Q_{DG_i} \leq Q_{DG_i}^{max} \quad (8)$$

Bus Voltage Limit [11].

The bus voltage magnitudes are to be saved inside appropriate working limits during the optimization technique. The *rms* value of the bus voltage involves only the fundamental component.

$$V_{min} \leq |V_i| \leq V_{max} \quad (9)$$

Where Vmin is the lower bound of bus voltage limit, Vmax is the upper bound of the voltage limits and $|V_i|$ is the root mean square (rms) value.

III. PROPOSED ALGORITHM

Many intelligent algorithms such as PSO, GSA, IGSA, GA, and others have applied optimal DG allocation in optimal allocating distributed generation. Optimization targets to determine the optimal placement for DG whilst they install in a

distribution network. This technique is primarily based on population-based search techniques that practice both random variation and selection. In this paper, the IGSA proposes determining the optimal placement and sizing of DG in the distribution system. The Newton Raphson load flow algorithm from MATPOWER is integrated into this optimization technique to obtain the minimum fitness functions for the total power losses, average THDV and the voltage profile of the system. The proposed technique uses to discover the excellent answer to the trouble in this paper. The optimization techniques considered in this study are PSO, GSA and IGSA.

A. Improved Gravitational Search Algorithm

The improved gravitational search algorithm (IGSA) proposes an optimization technique that improved from the original gravitational search algorithm [13]. Therefore, GSA needs to be improved to get a better search result applied in the electric distribution network. In the GSA concept, an agent of performance is considered by their masses since all the agents attract each other by the gravity force causes a global movement of all agents toward the agent of more massive masses [9]. Exploration and exploitation are two contradictory objectives that enhance the achievement of GSA successes [9]. However, the GSA's weaknesses were: first, the best agent is still exploring the global space even it was in the best position. The second weakness was that the best agent is still exploring the global space, even at position [18]. Improved Gravitational Search Algorithm (IGSA) presented to eliminate the weakness and improve the quality of the result and achieve the fastest convergence speed and global searchability. In the proposed IGSA, the chaotic dynamic applied to improve the searching behaviour and avoid premature convergence [19-20]. The flow chart of the IGSA algorithm is shown in Fig. 1. This method is applied in the 33-bus radial distribution system considering load variation.

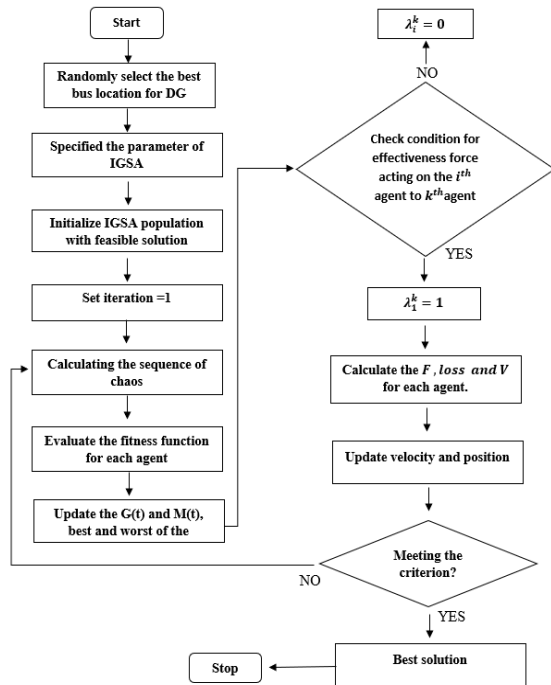


Fig. 1. IGSA Flow Chart.

B. Preliminary Assumption of the DGs

The data were obtained from bus radial system topology and used to illustrate the functionality of the proposed method to optimal placement and sizing of DG with considering load variation. The preliminary assumption of the DGs and all the parameters are set as follows:

- Voltage limit of the DGs: $0.90 \leq V \leq 1.05$.
- Power Generation limit of the DGs: $30\% \leq P \leq 60\%$ of the total connected load.
- Constant output generation with unit power factor is considered in the DG unit.
- Constant power output is modelled for all loads.
- The simulation is applied based on the changing load level.
- The installation of PVs solar units is same for each phase.
- The cost is not considered in this simulation.
- Renewable energy DG used, and only active power injected.

C. Case Studies

The control variable for this study was the size of DG, voltage deviation, and the location of DG. While based on the result obtained for entire real power base configuration for each variation of load and multi-objective function to minimize the losses, THD and voltage deviation, the simulation was done to find the best solution according to the case study.

The effect of the load variation is the main settings in this paper. The only supply source in the system is known as the slack bus has a constant voltage and phase angle. The maximum number of iterations is setting as 300 times for tuning process of each parameter. Six cases are considered in this paper regarding the impact of DG installed considering load variation toward power losses, average THDV and voltage deviation. All cases implemented on the Improved Gravitational Search Algorithm (IGSA) is proposed as an optimization strategy, and its overall performance is compared with different optimization techniques. Table I show the tabulated case studies for radial distribution system with a variety of load levels, present of DGs and optimization technique used.

TABLE I. CASE STUDIES

Case No.	Optimization Techniques			DGs Availability		
Case 1 (Load level 25%)	PSO	GSA	IGSA	No DG	1 DG	2 DGs
Case 2 (Load level 50%)						
Case 3 (Load level 75%)						
Case 4 (Load level 100%)						
Case 5 (Load level 125%)						
Case 6 (Load level 150%)						

IV. RESULTS AND DISCUSSION

The heuristic method that selects in this paper for optimal placement and sizing of DGs is tested on a 33-bus. The system loads are considered as a spot load, with the entire real power for base configuration is 3.72MW, 2.3 MVar with a real power loss of 0.203MW for the total connected load. The maximum iteration for the IGSA, PSO and GSA algorithm is set as 300. The supply source in this system at bus 1 is known as slack bus or reference bus with a fixed voltage and fed by a single source. The load is varying from 25% up to 150 %.

The foundation model of the framework has a single supply point with 33-buses and tie switches which are maintained generally open and is closed to vary circuit resistance for a reduction of losses or can be closed just during fault condition to support uninterrupted supply. The line diagram of the system is demonstrated in Fig. 2.

From the simulation of three optimization techniques, the result was obtained for the fitness function, total power losses, voltage deviation at various load level. Fig. 3 shows the convergence characteristic of among 30 simulation runs for three optimization techniques on a 100% load level with 1 DG installed in the 33-bus system. The result shows that the IGSA gives the best fitness compared to PSO and GSA according to total power losses, voltage deviation and average THDv.

Table II shows the base case results for power losses and the average voltage deviation for variation of load levels. By increasing the load level from 25% to 150%, it shows that the power losses increased while the average deviation decreased significantly.

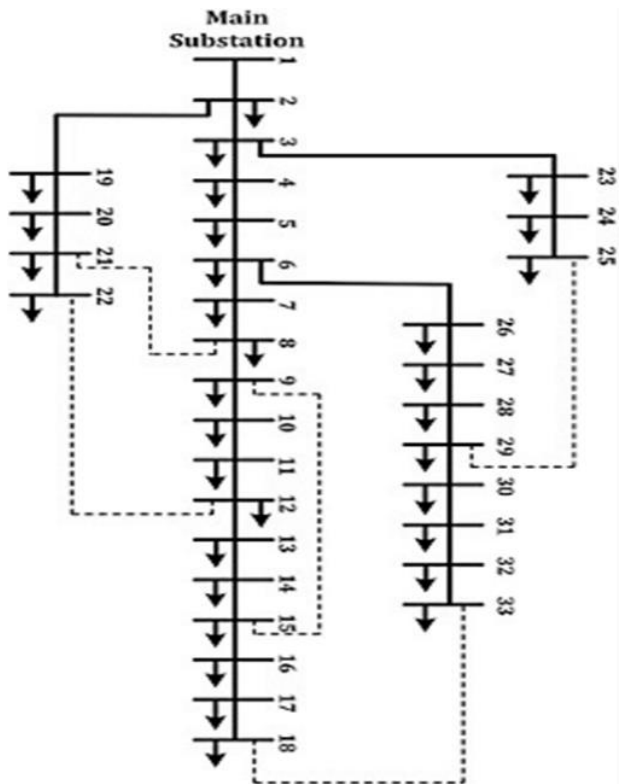


Fig. 2. IEEE 33-bus Distribution System.

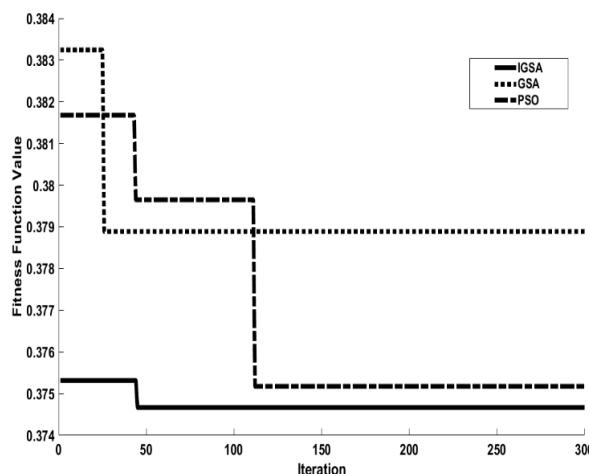


Fig. 3. Convergence Characteristic of GSA, PSO and IGSA Algorithms for 1 unit DG Installed in the 33-bus System.

TABLE II. BASE CASE FOR POWER LOSSES, AND AVERAGE VOLTAGE DEVIATION FOR VARIATION OF LOAD LEVEL

Variation of Load Level	DG availability	Losses(MW)	Average Voltage Deviation
Load 25%	No DG	0.011	0.987758
Load 50%		0.047	0.975303
Load 75%		0.11	0.962091
Load 100%		0.203	0.948484
Load 125%		0.33	0.934182
Load 150%		0.455	0.923545

Table III illustrates the best fitness values of various load level, the DG availability and the optimization techniques application. From the results in Table III, it shows that the IGSA technique gives the best fitness value compared to PSO and GSA for the various load level and the DG availability. Table III also shows the optimal DG location and the optimal DG size from the best fitness function. This result shows that the optimal DG size depends on the several factors such as the DG location, the variation of load level, and the power inject in the grid.

Table IV shows the DG impact on power losses for various load level and optimization techniques application. For 25% of load level, it shows that the losses are higher compared to the base case (without DG installed in the system). Therefore, the loss reduction becomes a negative value. However, for the rest of the cases (50% to 150% load level), the overall results show that the losses decreased significantly compared to the losses without DG. The IGSA shows the most excellent performance of the loss reduction for all cases compared to PSO and GSA.

Table V shows the DG impact on average voltage deviation for various load level and optimization techniques application. Overall, after DG installation in the system, the average voltage deviation had improved compared before DG installation. For different load level, the performance of the three optimization techniques is balanced. Its means that IGSA, not the only technique, give excellent performance in average voltage deviation.

TABLE III. THE BEST FITNESS VALUE OF THE DIFFERENT LOAD VARIATIONS, THE DG AVAILABILITY AND THE OPTIMIZATION TECHNIQUES

Variation of Load Level	DG availability	Techniques	Fitness	DG Location	DG Size
		PSO	0.75102	13	1.697
	1 DG	GSA	0.7496	3	1.6019
Load 25%		IGSA	0.74815	2	2.2676
		PSO	0.80135	8,16	2.9909, 3.1858
	2 DGs	GSA	0.79584	16,19	1.6032,3.0142
		IGSA	0.78449	30,33	2.9138,1.7892
		PSO	0.49282	28	1.9779
	1 DG	GSA	0.49398	2	2.4131
Load 50%		IGSA	0.49259	6	1.6006
		PSO	0.50676	9,20	3.0436,1.7591
	2 DGs	GSA	0.51018	6,29	3.1563,2.7326
		IGSA	0.50603	24,8	3.101,2.4787
		PSO	0.41026	33	2.53
	1 DG	GSA	0.41102	9	1.6022
Load 75%		IGSA	0.4103	12	2.5699
		PSO	0.42505	11,10	3.1852,1.8932
	2 DGs	GSA	0.43741	14,15	2.4907,2.8596
		IGSA	0.44133	27,3	3.0218,2.3309
		PSO	0.37518	16	1.9704
	1 DG	GSA	0.37873	16	3.2343
Load 100%		IGSA	0.37467	29	2.535
		PSO	0.38741	27,29	2.8227, 3.2842
	2 DGs	GSA	0.38895	26,28	2.6959,2.0149
		IGSA	0.38509	7,16	3.2642,2.9789
		PSO	0.35911	26	2.613
	1 DG	GSA	0.35938	30	1.6179
Load 125%		IGSA	0.35888	30	1.6025
		PSO	0.36079	22,16	2.0857,2.5866
	2 DGs	GSA	0.37882	6,5	3.0788,2.2627
		IGSA	0.35115	19,11	1.7974,3.0176
		PSO	0.36867	4	3.216
	1 DG	GSA	0.36907	27	2.9134

Load 150%		IGSA	0.36814	29	2.3019
		PSO	0.36699	17,6	2.5285,3.2116
	2 DGs	GSA	0.38354	10,8	2.3815,2.3805
		IGSA	0.36381	27,19	3.0054,1.7115

TABLE IV. DG IMPACT ON POWER LOSS FOR VARIATION OF LOAD LEVEL WITH THE APPLICATION OF THREE OPTIMIZATION TECHNIQUES USING THE 33-BUS SYSTEM

Variation of Load Level	DG availability	Technique	Losses (MW)	Losses reduction (%)
Load 25%	1 DG	PSO	0.013012	-18.29
		GSA	0.015509	-40.99
		IGSA	0.011094	-0.85
	2 DGs	PSO	0.077647	-605.88
		GSA	0.096913	-781.03
		IGSA	0.03915	-255.91
Load 50%	1 DG	PSO	0.019751	57.98
		GSA	0.019975	57.50
		IGSA	0.017331	63.13
	2 DGs	PSO	0.049554	-5.43
		GSA	0.071165	-51.41
		IGSA	0.048179	-2.51
Load 75%	1 DG	PSO	0.046412	57.81
		GSA	0.048765	55.67
		IGSA	0.043954	60.04
	2 DGs	PSO	0.077539	29.51
		GSA	0.089623	18.52
		IGSA	0.048416	55.99
Load 100%	1 DG	PSO	0.090213	55.56
		GSA	0.084135	58.55
		IGSA	0.064468	68.24
	2 DGs	PSO	0.11946	41.15
		GSA	0.129783	36.07
		IGSA	0.118329	41.71
Load 125%	1 DG	PSO	0.107128	67.54
		GSA	0.105893	67.91
		IGSA	0.105438	68.05
	2 DGs	PSO	0.135736	58.87
		GSA	0.158036	52.11
		IGSA	0.100804	69.45
Load 150%	1 DG	PSO	0.150441	66.94
		GSA	0.14558	68.00
		IGSA	0.142513	68.68
	2 DGs	PSO	0.156046	65.70
		GSA	0.143781	68.40
		IGSA	0.142916	68.59

TABLE V. DG IMPACT ON AVERAGE VOLTAGE DEVIATION FOR VARIATION OF LOAD LEVEL WITH THE APPLICATION OF THREE OPTIMIZATION TECHNIQUES USING THE 33-BUS SYSTEM

Load Variation	DG availability	Technique	Average Voltage Deviation	Voltage Deviation Improvement (%)
		PSO	0.989541	0.18
	1 DG	GSA	0.991644	0.39
Load 25%		IGSA	0.988804	0.11
		PSO	0.996333	0.87
	2 DGs	GSA	0.988579	0.08
		IGSA	0.99748	0.98
		PSO	0.989555	1.46
	1 DG	GSA	0.997005	2.23
Load 50%		IGSA	0.997455	2.27
		PSO	0.991949	1.71
	2 DGs	GSA	0.978855	0.36
		IGSA	0.991694	1.68
		PSO	0.990155	2.92
	1 DG	GSA	0.991082	3.01
Load 75%		IGSA	0.993249	3.24
		PSO	0.975883	1.43
	2 DGs	GSA	0.964536	0.25
		IGSA	0.994226	3.34
		PSO	0.981045	3.43
Load 100%	1 DG	GSA	0.984045	3.75
		IGSA	0.981877	3.52
		PSO	0.98877	4.25
	2 DGs	GSA	0.989192	4.29
		IGSA	0.991498	4.54
		PSO	0.970058	3.84
	1 DG	GSA	0.972759	4.13
Load 125%		IGSA	0.971385	3.98
		PSO	0.98505	5.45
	2 DGs	GSA	0.9797	4.87
		IGSA	0.987402	5.70
		PSO	0.961255	4.08
	1 DG	GSA	0.968894	4.91
Load 150%		IGSA	0.967659	4.78
		PSO	0.982023	6.33
	2 DGs	GSA	0.978175	5.92
		IGSA	0.977769	5.87

Table VI shows the DG impacts on average THDv for various load level and the optimization techniques. From Table VI, the increment of load level, the average THDv decreased significantly. However, the increment of the DG unit in the system is not significantly increased the average THDv in the system.

TABLE VI. DG IMPACT ON AVERAGE THDv FOR VARIATION OF LOAD LEVEL WITH THE APPLICATION OF THREE OPTIMIZATION TECHNIQUES USING THE 33-BUS SYSTEM

Variation of Load Level	DG availability	Technique	Average THDv (%)
Load 25%	1 DG	PSO	2.008568
		GSA	1.997636
		IGSA	1.993228
	2 DGs	PSO	2.146542
		GSA	2.197424
		IGSA	2.114125
Load 50%	1 DG	PSO	0.991635
		GSA	0.997575
		IGSA	0.993167
	2 DGs	PSO	1.013134
		GSA	0.994911
		IGSA	0.988376
Load 75%	1 DG	PSO	0.658226
		GSA	0.650975
		IGSA	0.644395
	2 DGs	PSO	0.664912
		GSA	0.662894
		IGSA	0.654173
Load 100%	1 DG	PSO	0.482678
		GSA	0.485075
		IGSA	0.478678
	2 DGs	PSO	0.489467
		GSA	0.49805
		IGSA	0.486432
Load 125%	1 DG	PSO	0.4043
		GSA	0.406037
		IGSA	0.403294
	2 DGs	PSO	0.394798
		GSA	0.439641
		IGSA	0.393116
Load 150%	1 DG	PSO	0.403448
		GSA	0.410749
		IGSA	0.402005
	2 DGs	PSO	0.407351
		GSA	0.483203
		IGSA	0.401549

After observation for losses in PSO, GSA and IGSA, the voltage profile is an observer for the proposed method on 100 % load in IGSA. In Fig. 4, the simulation result showed the voltage profile for a 100 % load level with IGSA, and the DG location is selected at buses 29 with 1 DG implementation. The voltage profile increases directly from base case 0.926pu up to 1.0038pu after DG was optimally installed in the distribution system guide based on IGSA method results; hence, when 2

DGs implement at the bus no. 7 and 16. The voltage profile tends to increase slightly to reference voltage magnitude within the specified limit. When two DGs are optimally installed, the voltage magnitude increased more efficiently, as shown in Fig. 4. The overall voltage profiles showed the increase in voltage magnitudes within the specified limit when the best allocation of DGs installed in the system. The comparison of the voltage profile when no DG installed with DG installed is illustrated in Fig. 4. Increasing of DG installed in the system drastically improved the voltage magnitudes but within the specified limit.

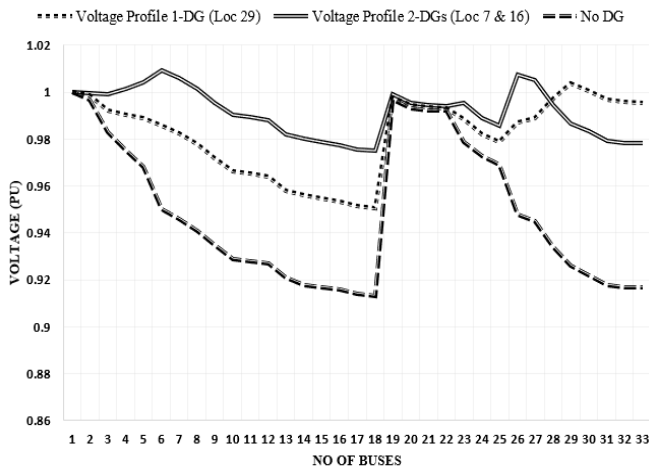


Fig. 4. IGSA Voltage Profile for 33 Bus System.

V. CONCLUSION

This paper has simulated the variable load in determining the optimal placement and optimal sizing of DG units using the IGSA technique. The multi-objective function was to minimize the total power, voltage deviation and THDv concerning the variation of load level. The results have shown that the proposed algorithm helps find a location to optimum total power loss of DGs in the power distribution system. The reduction of losses, voltage deviation and THDv has achieved after optimizing the DG placement and sizing. DG sizes are changing with a variety of load level and the location of DGs different for each load level. Thus, by considering the higher load level is much better for DG sizing in minimizing the losses, voltage deviation and THDv. It will reducing installation and maintenance costs as well.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka for all the support given. The paper's publication sponsored by the publication support scheme under the Centre for Research and Innovation UTeM (CRIM). Special thanks also go to the Office of the Assistance Vice-Chancellor (Development and Facilities Management), UTeM.

REFERENCES

[1] M.Syahzad Javid, T. Ma, J. Jurasz, J. Mikulik, "A hybrid method for scenario-based-techno-economic-environmental analysis of off-grid renewable energy system", *Renewable and Sustainable Energy Review*, Vol. 139, 110725, 2021.

[2] T. Ackermann and T. Ackermann, "Distributed Generation: A Definition Distributed generation: a definition," vol. 7796, 2016.

[3] S. S. Rawat, "Optimum Placement and Sizing of DGs Using Analytical Method for Different Types of Loads," pp. 1–5, 2015.

[4] Rau N.S., Wan Y.H., 1994 "Optimal location of resources in distributed planning" *IEEE. Trans. on Power Systems*, 9(1994), No. 1, 2014-2020

[5] D. Singh and R. K. Misra, "Effect of Load Models in Distributed Generation Planning," *Power Syst. IEEE Trans.*, vol. 22, no. 4, pp. 2204–2212, 2007.

[6] M. Abdel-Salam, M. T. El-Mohandes and L. Mahmoud, "A PSO-based Multi-objective Method for Optimal Weight Factors, Placement and Sizing of Multiple DG Units in a Distribution System," 2019 21st International Middle East Power Systems Conference (MEPCON), Cairo, Egypt, 2019, pp. 914-920, doi: 10.1109/MEPCON47431.2019.9008211.

[7] R. B. Magadam and D. B. Kulkarni, "Optimal Location and Sizing of Multiple DG for Efficient Operation of Power System," 2018 4th International Conference on Electrical Energy Systems (ICEES), Chennai, 2018, pp. 696-699, doi: 10.1109/ICEES.2018.8442393.

[8] P. Alinezhad, O. Z. Bakhoda, and M. B. Menhaj, "Optimal DG placement and capacity allocation using intelligent algorithms," 4th Iran. Jt. Congr. Fuzzy Intell. Syst. CFIS 2015, 2016.

[9] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "GSA: A Gravitational Search Algorithm," *Inf. Sci. (NY)*, vol. 179, no. 13, pp. 2232–2248, 2009.

[10] S. Kansal, B. B. R. Sai, B. Tyagi, and V. Kumar, "Optimal placement of distributed generation in distribution networks," *Int. J. Eng. Sci. Technol.*, vol. 3, no. 3, pp. 47–55, 2011.

[11] J. Kennedy, R. Eberhart, "Particle Swarm Optimization", *IEEE*, 1995, PP.1942.

[12] M. R. Ab Ghani, C. K. Gan, and I. J. Hasan, "Optimum Distributed Generation Allocation Using PSO to Reduce Losses and Voltage Improvement," 3rd IET Int. Conf. Clean Energy Technol. 2014, pp. 29 (6), 2014.

[13] A. F. A. Kadir, A. Mohamed, H. Shareef, A. A. Ibrahim, T., Khatib, and S. Energy, "An improved gravitational search algorithm for optimal placement and sizing of renewable distributed generation units in a distribution system for power quality enhancement," *J. Renew. Sustain. Energy*, vol. 6, no. 3, pp. 1–17, 2014.

[14] Barik, Soumyabrata, and Debapriya Das. "Determining the sizes of renewable DGs considering seasonal variation of generation and load and their impact on system load growth." *IET Renewable Power Generation* 12.10 (2017): 1101-1110.

[15] Essallah, Sirine, Adel Khedher, and Adel Bouallegue. "Integration of distributed generation in electrical grid: Optimal placement and sizing under different load conditions." *Computers & Electrical Engineering* 79 (2019): 106461.

[16] Peyman Karimyan, G.B. Gharehpetian, M. Abedi, A. Gavili, "Long term scheduling for optimal allocation and sizing of DG unit considering load variations and DG type", *International Journal of Electrical Power & Energy Systems*, Volume 54, 2014, Pages 277-287.

[17] H. N. Niri and A. Jalili, "Optimal placement of distributed generation in a power system by modified gravitational search algorithm," vol. 4, no. 1, pp. 30–38, 2016.

[18] A.A. Ibrahim, A. Mohamad, and A. Shareef, "A novel quantum-inspired the binary gravitational search algorithm to obtain optimal power quality monitor placement," *J. Appl. Sci* 12, pp. 882-830, 2012.

[19] M. Eslami et al., "An efficient particle swarm optimization technique with the chaotic sequence for optimal tuning and placement of PSS in power system," *Electr. Power Energy System*. 43, pp. 1467-1478, 2012.

[20] Lazzús, J. A., Vega-Jorquera, P., López-Carballo, C. H., Palma-Chilla, L., & Salfate, I., "Parameter estimation of a generalized Lotka–Volterra system using a modified PSO algorithm." *Applied Soft Computing*, 96, 106606, pp.1-7, 2020.

Resource Utilization Prediction in Cloud Computing using Hybrid Model

Anupama K C¹, Shivakumar B R², Nagaraja R³

Department of Information Science and Engineering
Bangalore Institute of Technology, Bangalore, Karnataka, India

Abstract—In cloud environment, maximum utilization of resource is possible with good resource management strategies. Workload prediction plays a vital role in estimating the actual resource required for successful execution of an application on cloud. Most of the existing works concentrated on predicting workloads which either showed clear seasonality/trend or for irregular workload patterns. This paper presents a new perspective in forecasting both seasonal and non-seasonal workloads. To accomplish this, a hybrid prediction model which is a combination of statistical and machine learning technique is proposed. Suppose the seasonality exists in the workload pattern, Seasonal Auto Regressive Integrated Moving Average (SARIMA) model is applied for prediction. For non-seasonal workloads Long Short-Term Memory networks (LSTM) or AutoRegressive Integrated Moving Average (ARIMA) model is used based on the results of normality test. This paper presents a prediction model which forecasts the actual resource required for diverse time intervals of daily, hourly and minutes utilization. The experimental results confirm that accuracy of the prediction of LSTM model outperformed ARIMA for irregular workload patterns. The SARIMA model accurately forecasts the resource usage for forthcoming days. This work actually helps the cloud service provider (CSP) to analyze the workload and predict accordingly to avoid over or under provisioning of the cloud resources.

Keywords—Workload prediction; SARIMA; LSTM; ARIMA; cloud service provider

I. INTRODUCTION

The Cloud computing is a utility computing model which is convenient to access the pool of computing resources such as physical machines, servers, applications, computing, storage, networks and various other services. The cloud computing model provides majorly three services such as software, platform and infrastructure as a service based on pay per usage model [1]. The elasticity feature of cloud computing enables the users to dynamically change the resource request periodically based on the demand. Due to fluctuating demands, the cloud manager must be able to leverage resources by provisioning and de-provisioning the resources to meet the current request. Insufficient provisioning of the resources causes Service Level Agreements (SLA) violation, poor Quality of Service (QoS), performance degradation which in turn causes customer dissatisfaction. On the contrary overprovisioning leads to wastage of resources which increases the cost and energy. For the seamless working of the system, judicious study of the dynamic and accurate resource provisioning is essential.

Workload prediction is one of the most critical and important aspect in managing cloud infrastructure in a flawless way. Every application requires resources to complete its execution and these resources come in the virtual form. The Cloud Data Centers (CDC) comprises of various resources like CPU, Memory, bandwidth, software, etc. which are allocated to the users on demand to complete their task execution. As per the previous works, it is well noted that resources provisioned to execute an application is always greater than actual resources required to complete it [2]. The reason for over provisioning of resources is to avoid SLA violations and to achieve QoS satisfaction. In most of the cases, the resources are being wasted in the process of allocation.

Accurate prediction can be used to decide the appropriate amount of resource to fulfill the demands. There is a need to employ a reliable and precise prediction model to achieve accurate estimation of the future workload. Usually in CDC, user's task arrives in an irregular pattern with heterogeneous resource requirement. This situation poses a major challenge to predict the precise workload [3]. Researchers have designed various workload prediction models and resource usage forecasting models, primarily concentrated on predicting CPU and memory utilization [4-7]. Various research works have used only statistical methods to predict workload and they are unable to predict accurate results for large and heterogeneous data [8]. Several research works have been carried out to address prediction of high dimensional and greatly varying cloud workloads using machine learning models. It is observed that, they were able to achieve promising prediction results. However, the statistical models are able to proactively predict temporal workloads in a controlled mode. Therefore, it is understood that combining both statistical and machine learning techniques when applied on heterogeneous data would result in better prediction accuracy [9]. Nevertheless, moderately a smaller number of research works has been carried out in the area of resource prediction at task level [10]. By predicting the resource utilization at task level aids in characterization of tasks, majorly impacts the process of task allocation, VM creation and capacity planning [11].

The main objective of this work is to accurately predict the CPU and memory utilization for different time intervals benefiting the cloud management for proper utilization of the available resources. The proposed Hybrid prediction model uses both statistical and machine learning approaches to achieve better quality prediction results with accuracy. This paper proposes a workload prediction model which is aimed to

predict the actual resource consumption of Central Processing Unit (CPU) and memory against provisioned resources. The pre-processed historical data is used to train the proposed prediction model. The predicted results are further used by the task classifiers to classify the tasks according to the resource utilization types which in-turn aids in resource management. Throughout this paper, prediction and forecasting has been used interchangeably. Remainder of the paper is structured as: Section 2 presents the overview of the existing works related to prediction using machine learning, statistical and hybrid methods in terms of resource utilization and accuracy. The Section 3 broadly explains proposed architecture and working principles of the prediction model. Section 4 presents the analysis of the workload trace, experimental setup and the obtained results. Finally, Section 5 concludes the work and mentions the future scope of the work.

II. RELATED WORKS

This section summarizes different prediction methods based on machine learning, statistical and hybrid approaches.

A. Machine Learning Methods

Machine learning and deep learning methods works on large and multivariate time series forecasting problems. Learning models outperforms when they are applied to complex and highly nonlinear data. They also make the most accurate long-term predictions. Chen et al. [1] applied L-PAW (deep Learning based Prediction Algorithm for cloud Workloads) utilized top-sparse encoder and gated recurrent unit block into Recurrent Neural Network (RNN) to achieve accurate prediction for high-dimensional workloads. Applying various evaluation metrics enhances in understanding the quality of the model. Yu et al. [6] developed a learning approach based on clustering to predict long-term workloads. In their work, K-medoid algorithm was used for clustering and multi-layer perceptron neural network for learning the patterns of workload and compared with non-clustering-based learning approach. Combining different prediction approaches can further improve accuracy of the results obtained. Furthermore, Qiu et al. [12] designed a deep learning prediction model for VM workload prediction using Deep Belief Network (DBN) with multiple-layered Restricted Boltzmann Machines (RBMs) and a regression layer. The authors have monitored only CPU usage and also the performance of the model is appreciable. Many more works have been proposed based on recurrent neural networks of deep learning. Wang et al. [13] and Guo et al. [14] proposed a prediction method using LSTM. The former proposed a model called LSTM_{tsw} to predict the future resource request trend of users and forecast CPU and memory resources and results proved that LSTM outperforms BPNN model. Whereas, the later worked on VM workload prediction using N-LSTM or the novel LSTM and compared the results with other LSTM variants in terms of prediction accuracy but the training and testing time required was high in their proposed model. Nguyen et al. [15] designed and implemented a new approach to predict workload by stacking Recurrent Neural Networks and Autoencoder on different datasets to compare prediction accuracy. Better prediction accuracy results may be possible if LSTM and autoencoder combination was used. It is understood that,

machine learning methods outperforms statistical method in terms of forecasting horizons and accuracy.

B. Statistical Methods

Statistical forecasting uses historical data to predict the future demands and these methods have been successfully used for short-term predictions. Calheiros et al. [8] presented cloud workload prediction for SaaS providers which was based on the ARIMA model to achieve the accuracy in resource utilization. Even though results showed an accuracy around 91%, there is a possibility to work on achieving better quality of service. Shyam et al. [16] proposed Bayesian model for accurate prediction of long-term and short-term resource requirements mainly considering CPU and memory intensive applications. Appreciable work has been done on predicting the resource utility. The understanding of the nature of workload can be further enriched if high-level metrics are used in prediction. Nashold et al. [17] forecasted CPU utilization in clusters using SARIMA and LSTM for both long term and short term tasks. The study concentrated only on predicting highest CPU utilization for the upcoming intervals of time but handling of memory utilization, which is as important as CPU is not considered in the work. Adhikari et al. [18] have proposed their work on time series models, salient features, related issues and importance of forecasting in various practical domain. Finally, it is observed that combining different prediction approaches can improve the forecasting accuracy. Parmezan et al. [19] has done an extensive study on evaluating statistical and machine learning model for time series prediction that deal with univariate data. They have experimented and compared the results using 55 real and 40 synthetic time series datasets. The work narrow downs the prediction by dealing with only univariate data, multivariate data are not considered in their work. From the overall study, it is found that statistical methods limits the understanding of forecasting horizons and accuracy.

C. Hybrid Prediction

The following researchers have attained better workload prediction accuracy by combining statistical, machine learning and mathematical models. Bi Jing et al. [3] proposed an approach that combines ARIMA and Haar wavelet to predict the number of tasks arriving at the successive time interval in the data center and results proved that hybrid approaches results in better prediction accuracy. Computing resource like CPU, memory etc., which plays a vital role in resource allocation are not addressed in the work. Janardhanan et al., [20] used LSTM and ARIMA models for forecasting CPU workloads in Google's cluster data. From the results it is found that LSTM has 20% less forecasting error when compared to ARIMA model and performed with better consistency. Experiment is implemented on a single machine and forecasted CPU utilization for that machine. Furthermore, Cetinski et al., [5] proposed an Advanced Model for Efficient Workload Prediction in the Cloud (AME-WPC) combining statistical and the machine-learning techniques to improve the accuracy of the workload prediction over time. The proposed approach proved to be efficient in terms of managing resource and minimizing the operational costs. But forecasting of specific cloud resource would aid in scaling of resource automatically and in the process of scheduling.

Islam et al., [9] presented an Error Correction Neural Network (ECNN) and Linear Regression learning algorithms for adaptive resource provisioning in cloud to forecast future resource demands especially for e-commerce applications. Predicting resource usage for different time intervals may be hourly, daily and monthly will provide a good insight for dynamically scaling the resource in cloud environment. Furthermore, Ullah et al. [21] designed a prediction model which takes real-time resource utilization and feeds these values to ARIMA and Autoregressive Neural Networks (AR-NN). The model is used to predict CPU utilization for the forthcoming four hundred minutes and it is observed that other physical resources are excluded in the study which are essential for predicting resource usage in IaaS cloud. From all these works, it is clear that combining both statistical and machine learning techniques when applied on heterogeneous data would result in better prediction accuracy.

III. SYSTEM MODEL

The proposed work forecasts the resource required for incoming tasks and predicting resource for the next given time period. This work considers CPU and Memory utilization, as they play an important role in minimizing cost and energy in CDC. This section presents an overview of the proposed prediction model as shown in the Fig. 1. Initially, the proposed models are trained and tested with the historical workload traces. The historical workloads trace comprises of the properties of executed tasks from the CDC. The historical data is fed to the pre-processing module for training and testing. The pre-processed module cleans the trace data and test the normality of the datasets. Afterwards, depending on the test results, the hybrid prediction module chooses the appropriate models for forecasting the CPU and memory usage. Secondly, depending on the test results, the hybrid prediction module chooses the appropriate models for forecasting the CPU and memory usage. Finally, the forecasted results are fed to Evaluation model, where the accuracy of the prediction is measured using statistical metrics.

Once the model is trained and tested with better accuracy, then the model is ready for the prediction of resources for the new incoming applications. Every user's applications are executed in the cloud platform using internet. Each application is divided as tasks for example, $T_1, T_2, T_3 \dots T_n$, etc. as shown in Fig. 1. The incoming tasks are fed to pre-processing module. The incoming job or task attribute values will undergo normality and seasonality test to choose the appropriate prediction model. The hybrid prediction model produces CPU and memory requirements to execute given task or application without the violation of SLA.

A. Pre-processing

The main goal of this process is to transform the obtained historical workload traces into a proper format as required by the proposed model. Since the available dataset/traces may contain noisy data, it is necessary to eliminate such before applying to any models. The cleaned data undergoes seasonality and normality test in the preprocessing module which checks normality for correlated data.

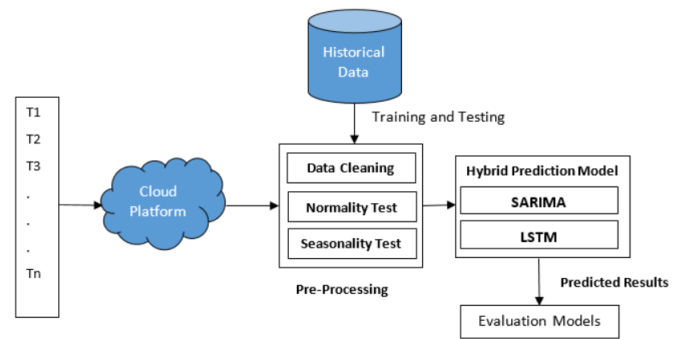


Fig. 1. Hybrid Resource Utilization Prediction Model.

Data Cleaning: The missing values, outlier values which are present in the trace data are cleaned by dropping such tuples from the dataset. The valuable information is retained and organized for further process. All the attributes do not contribute to the prediction process. Hence only those attributes which are very essential are extracted and stored separately. For example, in this experiment the attributes such as provisioned CPU, CPU used, provisioned memory, the actual memory used etc., are considered for CPU and memory utilization prediction. Thus, data cleaning process enhances the prediction accuracy.

Normality Test: The cleaned data has to undergo the normality test to determine, how likely the underlying dataset is normally distributed. Jarque-Bera test is applied to calculate kurtosis and skewness of the dataset [21]. If the test results show normal distribution then, statistical based prediction model is applicable else machine learning model is applied.

Seasonality Test: The simplest way to test and analyse for seasonality component such as daily, weekly or monthly in time series data by plotting the dataset. If the graph shows repeating spikes, that indicates there exists seasonality in the data. Hence for the data that exhibits seasonal patterns, SARIMA model is used for resource forecasting. On the other hand, non-seasonal time series data, ARIMA model is applied.

B. Hybrid Prediction Model

After exploring many works on prediction methods, it is understood that combining machine learning and statistical methods have proved to predict with better accuracy. Therefore, following two methods which are popularly known for time series prediction are used, namely:

- Seasonal ARIMA, which is an extension of ARIMA used in analyzing time series with seasonality.
- LSTM - a variant of RNN, which is capable of solving the problem of long-term dependency.

Since the work focuses on data containing trends and seasonality, SARIMA model was chosen which supports in analyzing the seasonal characteristics of the time series data. When it comes to forecasting data with complexity and non-linearity, LSTM method is applied, which is strong enough in identifying the complex pattern and structure in the given data. The following section discusses each of the method in detail.

1) **SARIMA Model:** Experiment uses SARIMA model which is similar to ARIMA with the exception that it takes seasonality into account. Seasonality is a regular pattern that appears in time series data, where changes are repeated over S time periods. Here S indicates the number of time periods till the pattern repeats again. For example, consider any monthly data appearing with high seasonality in a particular month and low seasonality in other months of the year. In such case, S=12 (months per year) and S=4 (for quarterly) is the span of the periodic seasonal behavior. Here, the time period of the day and month is considered.

In SARIMA both seasonal and non-seasonal factors are incorporated and denoted as SARIMA (p,d,q) × (P,D,Q)s. The lowercase notations denotes the non-seasonal component of the model, where p=non-seasonal AR (Auto Regression) parameter, d=non-seasonal differencing parameter, q= non-seasonal MA (Moving Average) parameter. The uppercase notations denotes the seasonal component of the model, where P=seasonal AR (Auto Regression) parameter, D=seasonal differencing states how many differencing orders to apply to make the time series stationary, Q= seasonal MA parameter and s= time span of repeating pattern [20]. The general form of the SARIMA model is given by (1):

$$\Phi_p(B^s) \phi(B) \nabla_s^D \nabla^d x_t = \Theta_Q(B^s) \theta(B) w_t \quad (1)$$

Where, $\{w_t\}$ is the Gaussian white noise process. s is the period of the time series. The AR and MA components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q. The seasonal AR and MA components are represented by $\Theta_P(B^s)$ and $\Theta_Q(B^s)$, and their orders are P and Q. Ordinary and seasonal difference components are indicated as ∇^d and ∇_s^D . (B) is the backshift operator and the expressions are presented from (2) - (7):

$$\text{AR: } \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad (2)$$

$$\text{Seasonal AR: } \Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \dots - \Phi_P B^{Ps} \quad (3)$$

$$\text{MA: } \theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q \quad (4)$$

$$\text{Seasonal MA: } \Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \dots + \Theta_Q B^{Qs} \quad (5)$$

$$\nabla^d = (1-B)^d \quad (6)$$

$$\nabla_s^D = (1-B^s)^D \quad (7)$$

The entire approach of SARIMA model is summarized as follows:

- Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are used to identify the seasonal and non-seasonal orders of SARIMA model.
- Coefficients of the model are estimated.
- Verify for the tolerance and predicting the future workload based on historical data.

2) **LSTM Model:** Recurrent neural networks are more popular and effective methods compared to traditional methods in machine learning. Along with stationary input and output patterns RNN also deals with sequences of arbitrary length. Since RNN suffer from short-term memory, LSTM

was generated as the solution to vanishing gradient & gradient explosion problems. This paper focuses on predicting the CPU and memory utilization requested by the users and experimented using LSTM [9], [20],[22]. In most of the recent works, LSTM has proved with better prediction accuracy when compared to other machine learning techniques especially for time series data.

LSTM operation is based on the mechanism of RNN. LSTM model is capable of capturing important features and remembering those information for a long interval of time. The Memory cell is a special feature of LSTM model which is an intermediate type of storage as in the Fig. 2. Memory cell is also called as gated cell as it is the one which decides about ignoring or preserving the memory information. Gated cells consist of sigmoid layer which outputs the numbers between zero and one. Value zero indicates to preserve information and one indicates to remove the information. The decisions are based on the weight values assigned during the training process. Hence the model learns about preserving what it needs and deleting the irrelevant information. Overall the LSTM model has three layers or gates: the forget gate, input gate and finally the output gate. The weights and the biases for the model are represented as : (W_f, W_i, W_g, W_o) and (b_f, b_i, b_g, b_o) .

Forget gate: Initial step in LSTM model is to decide which information needs to be removed from the memory. The forget gate or the sigmoid function looks at the values h_{t-1} and x_t to make this decision. The output f_t is a number between 0 and 1 indicating removing or preserving the information respectively. The output of this gate is calculated as in (8):

$$f_t = \sigma (W_f * [h_{t-1}, x_t] + b_f) \quad (8)$$

Where, b_f is a constant and it is called the bias value.

Input gate: This gate controls the flow of information by adding or deleting new information into the LSTM memory. The input gate has two parts: tanh and sigmoid layer respectively. The tanh layer creates g_t , a vector of new values that could be added to LSTM memory. The sigmoid layer i_t decides which values to be updated. Outputs of these two layers are computed as (9) and (10):

$$i_t = \sigma (W_i * [h_{t-1}, x_t] + b_i) \quad (9)$$

$$g_t = \tanh(W_g * [h_{t-1}, x_t] + b_g) \quad (10)$$

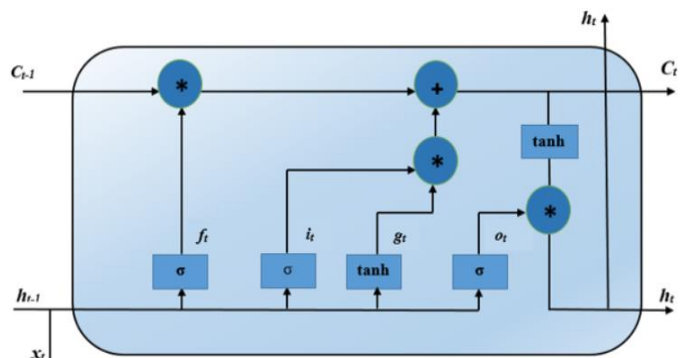


Fig. 2. Internal Structure of LSTM Block.

Combining these two layers provides an update to the LSTM memory. Updating the old value C_{t-1} into the new or current value C_t is computed by multiplying the old value by f_t the forget layer, and adding $i_t * g_t$. The mathematical equation is represented as (11):

$$C_t = f_t * C_{t-1} + i_t * g_t \quad (11)$$

Output gate: This primarily uses sigmoid layer to decide which part of LSTM memory will contribute to the final output. Later, a non-linear tanh function is performed to map values between -1 and 1. Lastly, the result of tanh is multiplied by the output of sigmoid layer. The output is calculated using the following (12) and (13):

$$o_t = \sigma (W_o * [h_{t-1}, x_t] + b_o) \quad (12)$$

$$h_t = o_t * \tanh(C_t) \quad (13)$$

Where o_t is the output value and h_t is the representation of the output and as value between -1 and 1.

C. Evaluation Criteria

To measure the accuracy of the work, two metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used to evaluate the errors of forecasting. MAE is a measurement metric where the absolute error is the absolute value of the difference between the forecasted value and the actual value and is calculated using (14).

$$MAE = \frac{1}{N} \sum_{i=1}^N (y_i^p - y_i^a) \quad (14)$$

MAPE is used in forecasting accuracy of a prediction model. Lesser MAPE value indicates better prediction accuracy in terms of percentage. It is defined as in (15).

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left(\frac{y_i^p - y_i^a}{y_i^a} \right) * 100\% \quad (15)$$

In the above formula, predicted value is y_i^p , the actual value is y_i^a and N is the number of the predicted values in the dataset.

IV. EXPERIMENTAL EVALUATION

This section explains the dataset and experimental results of the prediction model. The experiment is performed significantly on business-critical workload traces collected from Bitbrains cloud [23]. Bitbrains cloud is a mid-size datacenter majorly hosting business-critical workloads and mainly specialized in business computation and managed hosting for enterprises. The Bitbrains faststorage trace contains information of 1250 VMs connected to fast Storage Area Network devices. The dataset is available for 30 days duration with the sample rate of 5 minutes and organized as one file per VM. Each file contains seven performance metrics: the provisioned CPU capacity, the CPU utilization, the provisioned memory capacity, the actual memory utilized, the network I/O throughput, disk I/O throughput and the number of cores provisioned.

SARIMA: Bitbrains cloud data was used to check whether the dataset has seasonality component in it. Experiment was conducted by plotting 30 days data. The description about the

dataset is provided in the workload trace analysis part. After plotting the time series data, it is identified that there exists strong seasonality component in the dataset as in Fig. 3. It is observed that each data point looks similar to the data points of every other day. This interpretation leads to conclude that there is regularity in the patterns with respect to CPU and memory usage.

It is well known fact that ARIMA does not support seasonal component or it can be modelled for the non-seasonal time series data. ARIMA expects to remove (or reduce) seasonality by computing differences and the method is called differencing. Since there exists seasonality in the dataset SARIMA model is applied to predict the CPU and memory utilization for the upcoming days.

The first step is to identify the order of ARIMA model: AR (p), MA(q) and I(d) and it is done by plotting ACF and PACF at different lag lengths. The results of ACF and PACF functions clearly indicates the seasonal MA component and also to identify the maximum order of AR parameter estimation is done using Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) to select the best SARIMA model from other alternatives. Different parameter combinations was tried and ultimately choose the best model with lower AIC and BIC score. After estimating the parameters, the model was validated by testing against the actual Bitbrains dataset.

The forecasted CPU and memory values are tend to be close to the actual points. Fig. 4 and Fig. 5 shows actual and predicted graph of CPU and memory utilization, where black line indicates actual and red line indicates the predicted results.

The results are compared with the dataset of previous 29 days data. After selecting the best model, the forecasting of CPU and memory usage for the next three days are predicted. Various accuracy metrics like MAE and MAPE are used to test the predicted results.

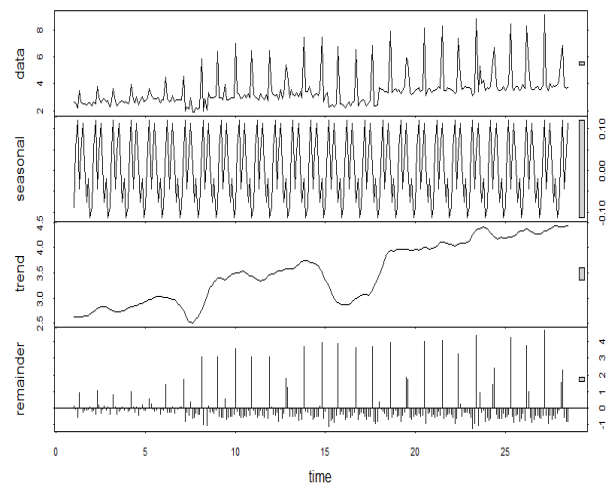


Fig. 3. Time Series Plot Representing Seasonal and Trend Components.

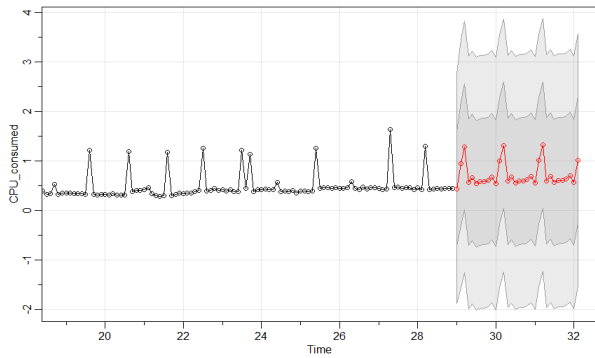


Fig. 4. CPU Utilization Prediction using SARIMA.

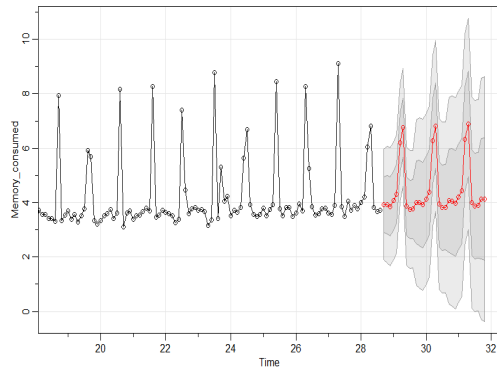


Fig. 5. Memory Utilization Prediction using SARIMA.

When experiment was compared with the predicted values of memory, it is observed that more than 10% of better prediction accuracy was found in case of CPU prediction. Since there is more fluctuation in the consumption of memory throughout the dataset, the values obtained after calculating MAE and MAPE are 6.08 and 1.52 respectively, whereas, in case of CPU usage prediction, the values of MAE were 5.83 and 0.49 for MAPE. Finally, it is found that the accuracy can be improved with the parameter estimation for the model. SARIMA model predicts the seasonality that occur over a period of time. Thus, SARIMA model is used to forecast the future resource consumption for the upcoming days and for months. This apparently helps the cloud service provider to analyze the trend and can avoid overprovisioning or underprovisioning of the resources.

LSTM: The implementation of LSTM model to evaluate the prediction accuracy of CPU and memory usage was efficiently tested on the Bitbrains dataset. The 30 samples of 5 minutes are average filtered to obtain a sample of 150 minutes. The CPU and memory forecasting are computed for both hourly and for minutes usage. Firstly, the experiment was conducted for predicting the CPU and memory consumption per hour interval. The .csv file consisting of CPU utilized and the actual memory utilized with respect to time was loaded into the working environment. Experiment was performed by varying number of training epochs and batch size. Epochs is the number of times the data is fed into the network and batch size allows to segment the data so that the network can process as small parts. It is found after the experiment that more the number of hidden layer size of LSTM, lesser was the accuracy. Thus, by modifying the weights and layers in the LSTM model, CPU and memory utilization prediction was accomplished. Fig. 6 represents the graph showing the actual and predicted values of CPU utilized in percentage vs Time in hours.

Similarly, the actual and predicted memory utilization in the units of 100MB with respect to time in hours is presented in Fig. 7. To illustrate the prediction accuracy, the LSTM model was compared with ARIMA model. LSTM model performed with better in terms of accuracy compared to ARIMA model. The predicted values of MAE and MAPE are shown in Table I.

Primarily the experiment focused on predicting CPU and memory considering per hour interval. Next focus was on predicting the same for every five minutes interval. The forecasted CPU and memory usage for every five minutes interval and for total of 24 hours is as shown in Fig. 8 and Fig. 9.

As observed in the Fig. 8 and 9, there is a high spike during 24th hour of the day. This infers that there is highest amount of CPU and memory is consumed at that particular time interval. Even then accuracy of the LSTM model performed extremely well when compared with ARIMA model. The calculated values of MAE and MAPE are shown in the Table I. MAE-Hr, MAPE-Hr and MAE-Min, MAPE-Min are the scores obtained for hourly and minutes CPU and memory usage prediction.

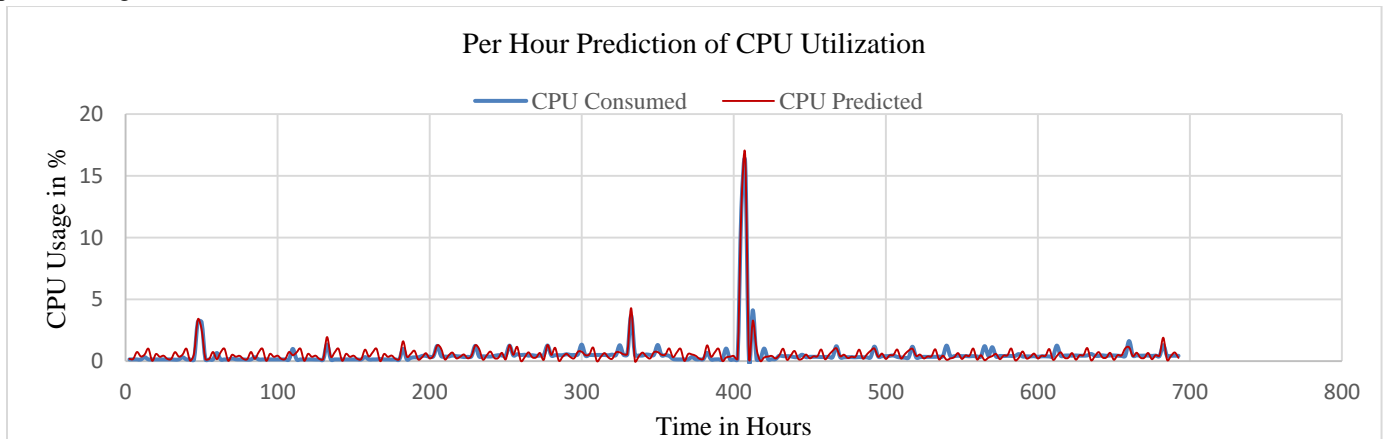


Fig. 6. Prediction of CPU utilization with LSTM Model - Time in Hours.

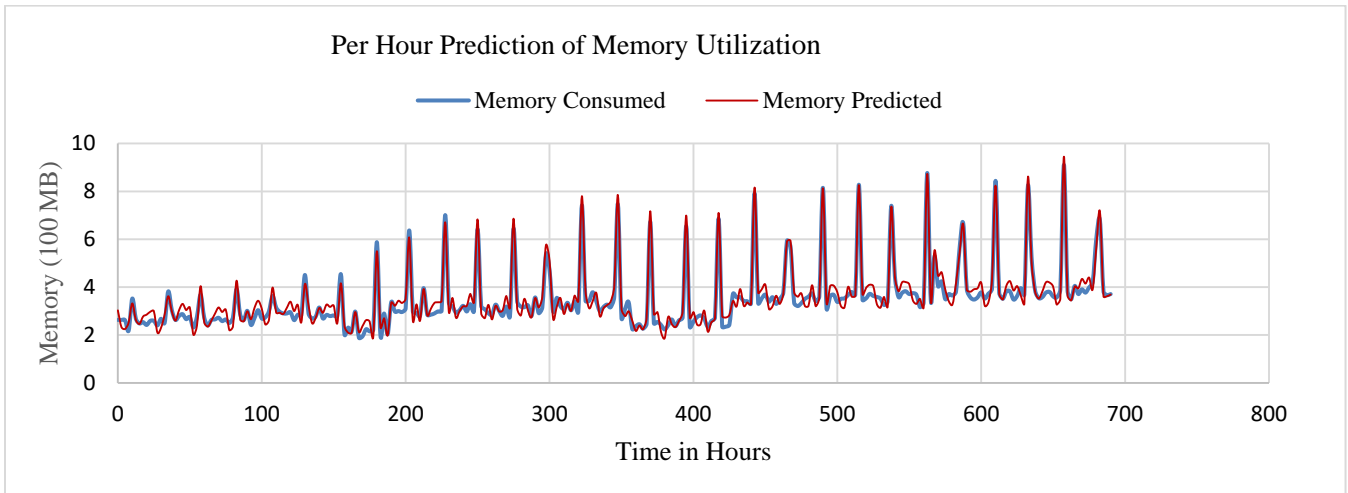


Fig. 7. Prediction of Memory utilization with LSTM Model - Time in Hours.

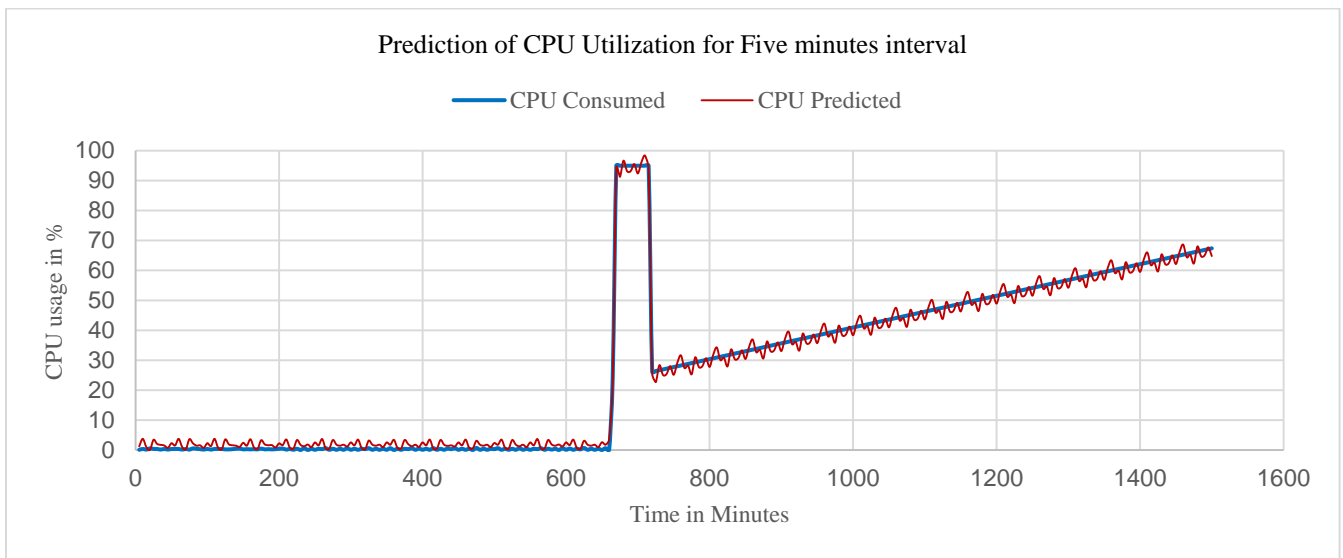


Fig. 8. Prediction of CPU utilization with LSTM Model - Time in Minutes.

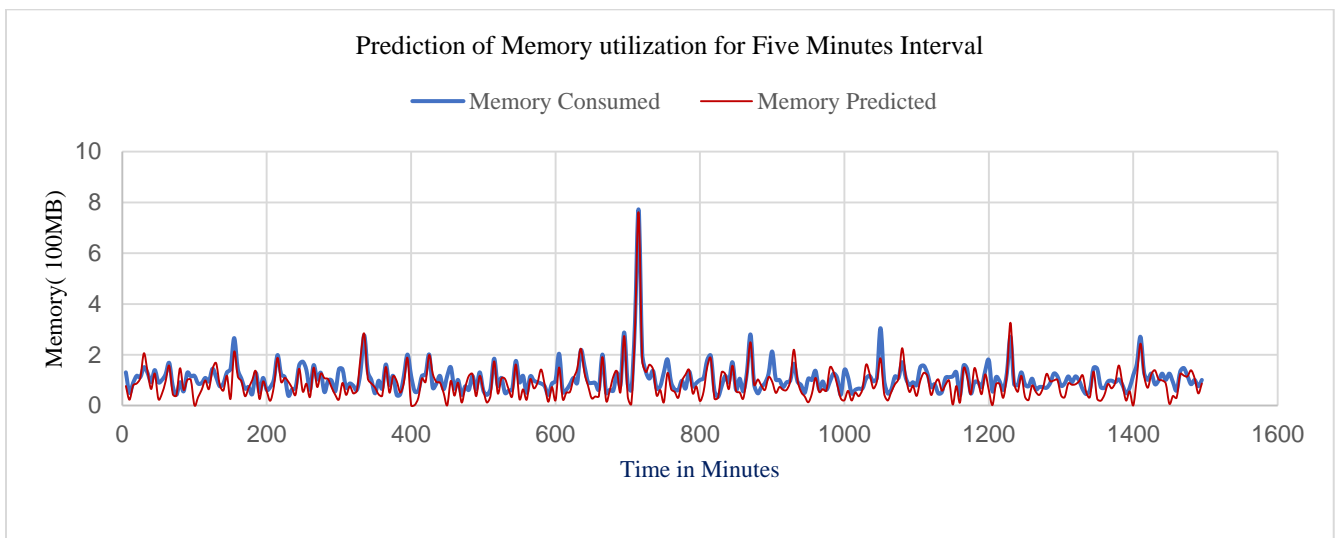


Fig. 9. Prediction of Memory utilization with LSTM Model - Time in Minutes.

TABLE I. PREDICTION ACCURACY COMPARISON - LSTM AND ARIMA MODELS

Model	MAE-Hour	MAPR-Hour	MAE-Minutes	MAPE-Minutes
LSTM-CPU	5.082	0.485	8.529	0.695
LSTM-Memory	6.3835	0.642	9.071	0.802
ARIMA-CPU	10.86	1.298	15.253	2.961
ARIMA-Memory	13.506	1.912	19.182	3.287

As seen in Table I, it is noticed that the prediction error was less during hourly forecasting of CPU and memory than forecasting for minutes. Results show that the LSTM model performs outstandingly well when compared to ARIMA model with reduced accuracy. ARIMA model performed better during hourly forecasting when compared to forecasting for minutes using the same model. Thus by overall observation it is found that LSTM models can selectively forget and retain the most relevant information as it flows through various layers. LSTM networks are the excellent model for forecasting workloads as it uses less computational resources when compared to RNN and more accurate forecasting results when compared to statistical models like ARIMA.

V. CONCLUSION

Accurate prediction of resource utilization is necessary for better resource management. This paper presented a prediction model for forecasting the resources like CPU and memory utilization. The model focuses on predicting both seasonality and random workload patterns. SARIMA model was able to predict the seasonality that occur over a period of three days for memory and CPU usage with a MAPE score of 1.52 and 0.49 respectively. Experiment was conducted using fastStorage, real trace data of Bitbrains data center. The results of the proposed method show that LSTM network has better prediction accuracy than ARIMA model. The model attained a MAPE score difference of 0.157 for hourly and 0.107 for minutes prediction of CPU and memory utilization. Thus the proposed workload prediction model is capable of predicting both seasonal and irregular workload patterns which aids in minimizing resource wastage. In future, study focuses on developing an approach particularly task level resource usage prediction.

REFERENCES

- [1] Chen, Zheyi, Jia Hu, Geyong Min, Albert Y. Zomaya, and Tarek El-Ghazawi. "Towards accurate prediction for high-dimensional and highly-variable cloud workloads with deep learning." *IEEE Transactions on Parallel and Distributed Systems* 31, no. 4 (2019): 923-934.
- [2] Liu, Jinwei, Haiying Shen, and Lihua Chen. "CORP: Cooperative opportunistic resource provisioning for short-lived jobs in cloud systems." In 2016 IEEE International Conference on Cluster Computing (CLUSTER), pp. 90-99. IEEE, 2016.
- [3] Bi, Jing, Libo Zhang, Haitao Yuan, and MengChu Zhou. "Hybrid task prediction based on wavelet decomposition and ARIMA model in cloud data center." In 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), pp. 1-6. IEEE, 2018.
- [4] Amiri, Maryam, and Leyli Mohammad-Khanli. "Survey on prediction models of applications for resources provisioning in cloud." *Journal of Network and Computer Applications* 82 (2017): pp. 93-113.
- [5] Cetinski, Katja, and Matjaz B. Juric. "AME-WPC: Advanced model for efficient workload prediction in the cloud." *Journal of Network and Computer Applications* 55 (2015): pp. 191-201.
- [6] Yu, Yongjia, Vasu Jindal, Farokh Bastani, Fang Li, and I-Ling Yen. "Improving the smartness of cloud management via machine learning based workload prediction." *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2, pp. 38-44. IEEE, 2018.
- [7] Kaur, Gurleen, Anju Bala, and Inderveer Chana. "An intelligent regressive ensemble approach for predicting resource usage in cloud computing." *Journal of Parallel and Distributed Computing* 123 (2019): pp. 1-12.
- [8] Calheiros, Rodrigo N., Enayat Masoumi, Rajiv Ranjan, and Rajkumar Buyya. "Workload prediction using ARIMA model and its impact on cloud applications' QoS." *IEEE transactions on cloud computing* 3, no. 4 (2014): pp. 449-458.
- [9] Islam, Sadeka, Jacky Keung, Kevin Lee, and Anna Liu. "Empirical prediction models for adaptive resource provisioning in the cloud." *Future Generation Computer Systems* 28, no. 1 (2012): pp. 155-162.
- [10] Borkowski, Michael, Stefan Schulte, and Christoph Hochreiner. "Predicting cloud resource utilization." In *Proceedings of the 9th International Conference on Utility and Cloud Computing*, pp. 37-42. 2016.
- [11] Anupama, K. C., R. Nagaraja, and M. Jaiganesh. "A Perspective view of Resource-based Capacity planning in Cloud computing." *1st International Conference on Advances in Information Technology (ICAIT)*, pp. 358-363. IEEE, 2019.
- [12] Qiu, Feng, Bin Zhang, and Jun Guo. "A deep learning approach for VM workload prediction in the cloud." *17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 319-324. IEEE, 2016.
- [13] Wang, Hengjian, John Pannereselvam, Lu Liu, Yao Lu, Xiaojun Zhai, and Haider Ali. "Cloud workload analytics for real-time prediction of user request patterns." *IEEE 20th International Conference on High Performance Computing and Communications*; pp. 1677-1684. IEEE, 2018.
- [14] Guo, Wei, et al. "Short-Term Load Forecasting of Virtual Machines Based on Improved Neural Network." *IEEE Access* 7 (2019): pp. 121037-121045.
- [15] Nguyen, Hoang Minh, Sungpil Woo, Janggwan Im, Taejoon Jun, and Daeyoung Kim. "A workload prediction approach using models stacking based on recurrent neural network and autoencoder." *IEEE 18th International Conference on High Performance Computing and Communications* pp. 929-936. IEEE, 2016.
- [16] Shyam, Gopal Kirshna, and Sunilkumar S. Manvi. "Virtual resource prediction in cloud environment: a Bayesian approach." *Journal of Network and Computer Applications* 65 (2016): pp. 144-154.
- [17] Nashold, Langston, and Rayan Krishnan. "Using LSTM and SARIMA Models to Forecast Cluster CPU Usage." *arXiv preprint arXiv:2007.08092*, 2020.
- [18] Adhikari, Ratnadip, and Ramesh K. Agrawal. "An introductory study on time series modeling and forecasting." *arXiv preprint arXiv:1302.6613*, 2013.
- [19] Parmezan, Antonio Rafael Sabino, Vinicius MA Souza, and Gustavo EAPA Batista. "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model." *Information sciences* 484 (2019): pp. 302-337.
- [20] Janardhanan, Deepak, and Enda Barrett. "CPU workload forecasting of machines in data centers using LSTM recurrent neural networks and ARIMA models." In *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)*, pp. 55-60. IEEE, 2017.

- [21] Zia Ullah, Qazi, Shahzad Hassan, and Gul Muhammad Khan. "Adaptive resource utilization prediction system for infrastructure as a service cloud." *Computational intelligence and neuroscience* 2017.
- [22] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): pp. 1735-1780.
- [23] Iosup, Alexandru, Hui Li, Mathieu Jan, Shanny Anoep, Catalin Dumitrescu, Lex Wolters, and Dick HJ Epema. "The grid workloads archive." *Future Generation Computer Systems* 24, no. 7 (2008): pp. 672-686. Available: <http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains>.

An Experiment for Outdoor GPS Localization Enhancement using Kalman Filter with Multiantenna Consumer-Grade Sensors

Phudinan Singkhamfu¹, Parinya Suwansrikham²
College of Arts, Media and Technology
Chiang Mai University, Chiang Mai
Thailand

Abstract—Consumer-Grade global positioning system (GPS) is widely used in many domains. The obvious issue of this consumer-grade device is low accuracy and reading fluctuation results. In terms of using an application that requires a more precise location, the output could be difficult. In this study, the authors deploy various methods to reduce the global positioning system data fluctuation and present field test results. Two main types of the device worked together to collect data from global positioning systems, such as Microcontroller for algorithm processing and presenting data and global positioning system receivers for receiving data from a satellite. We combine three global positioning system modules to received signals in a single device and test calculated data compared with the Kalman filtering methods in many cases, including moving and static devices. Implementing the Standard Kalman Filter to multiple global positioning system Modules has improved the constancy of cheap global positioning system equipment. The experiment algorithm is presented significant improvement to overcome the retrieved data fluctuation problem. This study's contribution will enable creating a cheap global positioning system locator device for various applications that require more accuracy than the standard consumer-grade receiver.

Keywords—Global positioning systems accuracy; Kalman; multi global positioning systems; global positioning systems pointer; global positioning systems enhance; filtering algorithm

I. INTRODUCTION

It is widely known that Global Positioning System or GPS [1], which was invented during the 1960s–1970s, has been broadly used in several sectors such as service, academics, economics, and development. It can safely be said that GPS is a fundamental technology commonly found in our daily lives.

Even though the positioning system of GPS is relatively new and has been further developed into numerous inventions in the past five decades, it does not particularly mean that GPS is the most accurate system, especially when compared to GNSS (Global Navigation Satellite Systems), which is a more expensive specialized navigation system [2,3].

Although, a consumer-grade GPS is less accurate, and current computer technology can improve its precision with algorithm commands. Kalman Filter is an algorithm used to estimate possible variables and lower the discrepancy of GPS. In consequence, it is making the inexpensive GPS locator for many projects that limited fund is complicated, for example,

the guidance device in entree level drone, personal location device, and forest fire locator for the rescue team.

However, there can still be an unsatisfying discrepancy if Kalman Filter is solely applied to just one device [4]. On the other hand, if several GPS devices are integrated with Kalman Filter to determine a more reliable statistical means, the results can be more efficient compared to using only one GPS device [5,6]. The prototype also has the limitation of hardware durability due to using a prototype grade sensor and Universal printed circuit board (PCB).

This experiment aspires to present a new concept derives from combining two calculation techniques using different algorithms but sharing the same objectives. This innovation can elevate the efficiency of the system using only one of the calculation techniques. It is expected that this innovation is an alternative to better technological development.

II. BACKGROUND

For technological development, consumer-grade smart devices typically contained parts or sensors that could easily be found in the market due to cheap costs and accessibility while still generating acceptable precision. For example, a Quadcopter drone could solely control the Hover Control System by itself using the Microcontroller and Inertial Measurement Unit (IMU), which could be found in general markets [7].

Lower prices and convenient accessibility came with lower efficiency compared with other more expensive specialized devices. Moreover, there have been many times that the instability of the devices results in inaccuracy. One of the most encountered problems was the instability of GPS in navigating and positioning. The accuracy of 95% of the reviewed literature was approximately 10 – 15 meters from the designated location, both Latitude and Longitude [8]. This was since several environmental factors were affecting the accuracy of the results of consumer-grade GPS devices; for example, there was a Doppler Shift phenomenon where the increased speed of GPS devices generated very low discrepancy [9], and the weather during a clear sky generated 0 – 2 meter discrepancy, while during a closed canopy condition, the discrepancy could be up to 9 meters [10].

B. Standard Kalman Filter Approach

Kalman Filter is a set of computer commands used to predict possible outcomes of linear equations based on estimations from Mean Square Error from historical data.

This experiment also used Kalman Filter with the same algorithm as the previous study [21], which consisted of three steps: 1) Initialization: initialed the variables used in prediction, 2) Prediction: calculated data for the possible outcomes, and 3) Update: currently collected data for the prediction of the next set of data. Prediction and Update functioned together recursively for data prediction using Kalman Gain as variables determining the future's possible outcomes would be according to the current data. The equations for Standard Kalman Filter Data Prediction were:

$$\text{Initial } \hat{x}_{k-1} \text{ and } P_{k-1} \quad (1)$$

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k \quad (2)$$

$$P_k^- = AP_{k-1}A^T + Q \quad (3)$$

$$K_k = P_k^- H^t (HP_k^- H^t + R)^{-1} \quad (4)$$

$$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-) \quad (5)$$

$$P_k = (1 - K_k H)P_k^- \quad (6)$$

The objective of these equations was to find an estimated value of the data at time K, aka \hat{x}_k , based on data collected from the current time (z_k) according to K_k (Kalman Gain), This was a crucial variable that varied directly with data from the past (1). Equation (2) and (3) were Prediction State where roughly estimated data were stored in \hat{x}_k^- (Prior estimate) and P_k^- (Prior error covariance) before being used later in Update, which was related to (4), (5), and (6) to finally yielding results in the estimate called \hat{x}_k .

IV. EXPERIMENT

Upon turning on the signal receivers, the calibration took 30 seconds before data collection started. Every collected GPS data would be displayed on the Serial Monitor of Arduino IDE. The data collection lasted at least 30 minutes, timed by a time switch. Every read needed approximately 2 – 3 seconds, and after which, all collected data were stored for further calculation.

The first experiment was to collect GPS data while the sensors were completely still. This experiment's location was the open space near the reservoir with no high buildings within a 100-meter radius from the receivers' position. The experiment was conducted at around 5.30 pm, during a clear sky with no visible cloud. The total time spent was 33 minutes and 2 seconds.

The following experiment was to collect GPS data while the sensors were continually moving. The location for this experiment was in the city, surrounded by no higher than 4-story buildings. The experiment was conducted at around 5.32 pm, during a clear sky with no visible cloud. The GPS receivers were sticking out from a backpack while the backpack carrier walked for 2.76km with the average speed at 7 – 8 m/hr, referring to Nike Run Club Application. The total time spent was 41 minutes and 30 seconds.

There were two rounds of data collection, one when the sensors were completely still and the other when the sensors were continually moving. For the one when the sensors were completely still, there were 663 sets of data collected, while for the one when the sensors were.

The picture on the left of Fig. 2 showed the location of 663 sets of GPS data read from all three sensors with no movement after visualizing on Grafana Application. This data collection lasted 1,982 seconds, or 33 minutes and 2 seconds. On average, each read took 2.99 seconds.

The GPS data read from each of the moving sensors was shown in the right picture of Fig. 2. This data collection contained 839 sets of GPS data and lasted 2,326 seconds, 3 minutes, and 46 seconds. On average, each read took 2.77 seconds.

The altitude above sea level read when the sensors were completely still and when they were constantly moving were represented by red, green and blue lines, respectively. On the left of Fig. 3, the range of the sensors with no movement was at 16.5 meters, with the lowest at 316.7 meters and the highest at 343.2 meters. Meanwhile, the range of the moving sensors was at 39.8 meters, with the lowest at 267.6 meters and the highest at 307.4 meters, as shown on the right of Fig. 3.

The collected data would then be calculated for the distribution of data, namely the Maximum, Minimum, Range, Standard Deviation, Mean Deviation, and Variance, derived from each sensor.

The distribution of Latitude and Longitude were shown in Table II, and True Altitude (Altitude Above Sea Level) was shown in Table III. However, the experiment with moving sensors was not calculated for the distribution of data because the actual position was changed continuously, meaning that all data could not be used to find the current location's distribution.

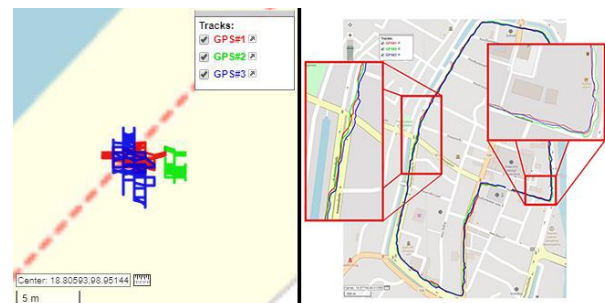


Fig. 2. GPS Positioning of all Three Sensors when the Equipment was Entirely Still (Left) and when the Equipment was Constantly Moving (Right).

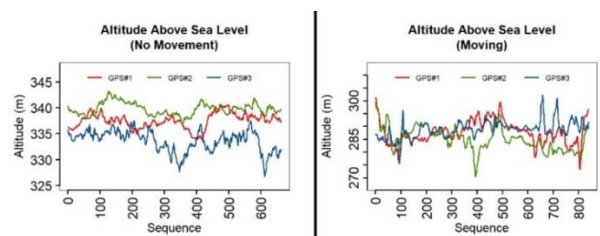


Fig. 3. Altitude above Sea Level Received from all Three Sensors when the Equipment was Completely Still (Left) and when the Equipment was Moving (Right).

TABLE II. STATISTICAL DATA OF LATITUDE AND LONGITUDE MEASURED FROM MULTIPLE GPS RECEIVERS WHILE THE RECEIVERS WERE COMPLETELY STILL

	GPS#1		GPS#2		GPS#3	
	LAT	LNG	LAT	LNG	LAT	LNG
Max	18.805810	98.951019	18.805807	98.951034	18.805820	98.951011
Min	18.805789	98.950973	18.805782	98.951019	18.805765	98.950981
Range (m)	2.34	5.12	2.78	1.67	6.12	3.34
S.D.	0.0000046	0.0000088	0.0000049	0.0000041	0.0000119	0.0000064
Variance	2.08E-11	7.79E-11	2.39E-11	1.70E-11	1.43E-10	4.05E-11

TABLE III. STATISTICAL DATA OF ALTITUDE ABOVE SEA LEVEL FROM THREE RECEIVERS IN BOTH NON-MOVING AND MOVING CONDITIONS (METER)

	No Movement			Moving		
	GPS#1	GPS#2	GPS#3	GPS#1	GPS#2	GPS#3
Max	340.70	343.20	337.50	301.30	299.30	307.40
Min	333.80	336.90	326.70	270.00	267.60	272.30
Range	6.90	6.30	10.80	31.30	31.70	35.10
S.D.	1.47	1.19	2.04	4.65	4.63	3.73
Variance	2.17	1.41	4.17	21.64	21.46	13.89

Latitude, Longitude and True Altitude were calculated to improve the stability of data using six methods. From all of the six methods, there were three interesting methods when applied to all ten scenarios as presented here.

A. Implement Kalman Filter to Data at a Certain Time and then Measure the Averages

This method conducted two calculations. The first calculation was implementing Kalman Filter to data from each sensor since it was found in previous studies that Kalman Filter could lower discrepancy to a certain level. However, the results were not efficient enough to stabilize the data [11]. Therefore, the second calculation for this method aimed to elevate the data improvement by measuring the averages using (9).

$$Lat_{avg} = \frac{\sum_{i=1}^n C_{lat_i}}{n} \tag{1}$$

$$Lng_{avg} = \frac{\sum_{i=1}^n C_{lng_i}}{n} \tag{8}$$

$$Alt_{avg} = \frac{\sum_{i=1}^n C_{alt_i}}{n} \tag{9}$$

The results from (7), (8), and (9) were the GPS locations and altitudes when the sensors were completely still as shown in Fig. 4.

The purple area was the one where Kalman Filter was implemented before measuring the averages. It was noticeable that the area was narrower compared to the other three sets of unprocessed data from three sensors due to the decreased data distribution. Table IV showed the statistical data with significantly decreased deviation compared with unprocessed data in Table II and Table III.

B. Measure the Averages, and then Implement Kalman Filter

This method was similar to the first method, and the difference was only that each set of data from all three

receivers were used to calculate the averages before implementing Kalman Filter.

Even though the Ranges of latitude, Longitude, and True Attitude of this method were the same as the first method, this method's statistical variance was significantly lower. As shown in Fig. 5, the second method's data distribution was remarkably similar to that of the first method, making it hard to distinguish via observation. From Table V, the variances of the GPS positions of both methods were slightly different, while the altitudes bore no difference at all at two decimal places.

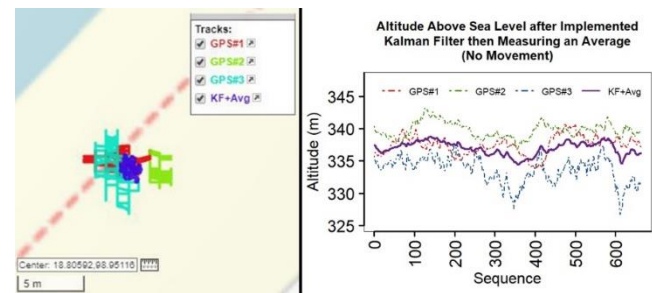


Fig. 4. GPS Positions and Altitude above sea Level Received from all 3 GPS Receivers after Implementing Kalman Filter and then Measuring the Averages when the Equipment was Completely Still.

TABLE IV. STATISTICAL DATA OF LATITUDE, LONGITUDE, AND ALTITUDE ABOVE SEA LEVEL FROM THREE RECEIVERS AFTER IMPLEMENTING KALMAN FILTER AND MEASURING THE AVERAGES WHILE THE SENSORS WERE COMPLETELY STILL

	$(Lat_k)_{avg}$	$(Lng_k)_{avg}$	$(Alt_k)_{avg}$
Max	18.805804	98.951013	338.76
Min	18.805785	98.950998	334.44
Range (m)	2.11	1.67	4.32
S.D.	0.0000041	0.0000035	1.01
Variance	1.72252E-11	1.24139E-11	1.02

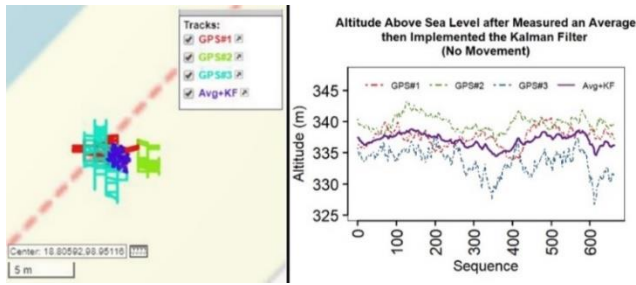


Fig. 5. GPS Positions and Altitude above Sea Level from all 3 GPS Receivers after Measuring the Averages and then Implementing Kalman Filter when the Equipments were Completely Still.

TABLE V. STATISTICAL DATA OF LATITUDE, LONGITUDE AND TRUE ALTITUDE AFTER MEASURING THE AVERAGES AND THEN IMPLEMENTING KALMAN FILTER WHILE THE SENSORS WERE COMPLETELY STILL

	$(Lat_k)_{avg}$	$(Lng_k)_{avg}$	$(Alt_k)_{avg}$
Max	18.805804	98.951013	338.76
Min	18.805785	98.950997	334.44
Range (m)	2.11	1.78	4.32
S.D.	0.0000041	0.0000035	1.01
Variance	1.70901E-11	1.25134E-11	1.02

C. Use the Data to Measure the Averages and then Implementing Kalman Filter for Every 2 to N Terms

It was found that using GPS data to calculate for averages before implementing Kalman Filter yielded better results; therefore, for this third method, every N Term was measured for averages before Kalman Filter was implemented, N being the Interval Number of data calculated for averages. For example, if $N = 3$, the system would read GPS data 3 times and then used these three values for calculation. The average gained from each GPS receiver were then added together and divided by the number of receivers (3 in this particular case) to find the average of Multiple Sensors, which were then implemented with Kalman Filter. This method aimed to observe the tendency of data in the case that the Interval of finding averages kept increasing while the GPS receivers bore no movement.

Table VI found that data distribution tended to keep decreasing when N (average Interval) increased. Upon checking the range of distribution, when Interval equaled 2, 3, and kept going to the total number (N), it could be seen that for every increasing N, the standard deviation decreased and tended to keep decreasing. The change of the graph's trend was noticeable in Fig. 6, which showed the comparison of data

TABLE VI. STATISTICAL DATA OF LATITUDE, LONGITUDE, AND ALTITUDE AFTER MEASURING FOR THE AVERAGES OF EVERY N TERM INTERVAL THEN IMPLEMENTING KALMAN FILTER WHILE SENSORS BORE NO MOVEMENT

	Measurement Interval = 2 terms			Measurement Interval = 3 terms			Measurement Interval = 1, 2, ..., 663terms		
	LAT	LNG	ALT	LAT	LNG	ALT	LAT	LNG	ALT
Max	18.805803	98.951012	338.66	18.805802	98.951011	338.58	18.805803	98.951006	337.62
Min	18.805786	98.950998	334.58	18.805786	98.950998	334.72	18.805796	98.951001	336.56
Range (m)	1.89	1.56	4.08	1.78	1.45	3.86	0.78	0.55	1.06
S.D.	0.0000040	0.0000033	0.98	0.0000038	0.0000031	0.95	0.0000011	0.0000011	0.29
Variance	1.57E-11	1.07E-11	0.96	1.43E-11	9.48E-12	0.91	1.28E-12	1.21E-12	0.08

calculated with this method with average calculation at 2, 3 intervals and from 1 to 663 terms.

For the case that the equipment was constantly moving, this method calculating for averages from 1, 2, to 663 terms would not be used. This was due to the fact that, when the Interval of the averages were increased, the data would start moving towards the center of the data as shown in Fig. 7 where the path of data at Interval 1 to the total number at 663 sets for the case that the equipment was continually moving. The Purple Line and the Blue line represented calculations with both Kalman Filter and Average Measurement. It is evident that when time passed, the path was compressed towards the center of the data. Therefore, this method was not used for moving equipment.

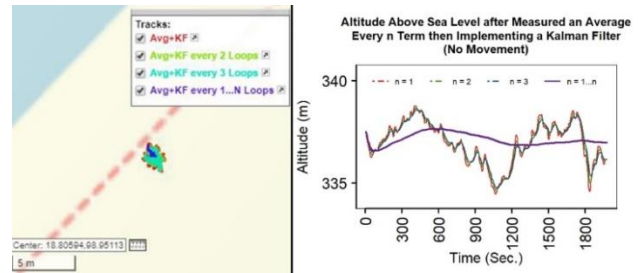


Fig. 6. GPS Positions and Altitude above Sea Level from all 3 GPS Receivers after Measuring the Averages of Every n Term Interval then Implementing Kalman Filter while Sensors bore no Movement.



Fig. 7. GPS Positions after all Three Sensors were Measured for Averages and Implemented with Kalman Filter Every N Term, and after Data from all Three Sensors were Implemented with Kalman Filter and then Measured for Averages in the Case that the Equipment was Continually Moving.

V. RESULT ANALYSIS

This experiment demonstrated three methods used to lower the distribution of data: 1) Implementing Kalman Filter, 2) Finding Averages and 3) Implementing Kalman Filter and Finding Averages. From the distribution data shown in Table VII, the least effective method was solely implementing Kalman Filter. The Standard Deviations of GPS Coordinates (Latitude, Longitude) while the GPS receivers were completely still were decreased by 3.39% and 10.7% on average; whereas the Standard Deviations of True Altitude were decreased by 2.99% and 6.23% for the non-moving equipment and the moving equipment, respectively. Even though it is evident that Kalman Filter could help reduce the distribution of data, but its efficiency was too low to be used with projects which needed data stabilization, as previously mentioned in earlier studies regarding Kalman Filter [11].

Next up was the method where data were used to find averages. This method significantly increased the stability and decreased data distribution better than the one implementing only Kalman Filter. The Standard Deviation of Latitude and Longitude for non-moving equipment were decreased by 49% and 75.21%. The Standard Deviation of True Altitude for non-moving equipment and moving equipment was decreased by 65.45% and 40.94%, respectively. They were resulting in more stability compared to the method solely implementing Kalman Filter.

Implementing both Kalman Filter and average measuring to improve data stability could be further divided into four sub-methods: 1) They were using results after implementing Kalman Filter to find averages, 2) using Kalman Filter results to find averages of every data from 1 to N loop, 3) using data after finding averages to implement Kalman Filter, and 4) using the averages of every data from 1 to N loop to implement Kalman Filter. Based on all these sub-methods statistical data, it was found that implementing Kalman Filter and average findings could better stabilize the data compared to applying

only one method. From Table VII, in the case that Kalman Filter was implemented before average measuring with non-moving equipment, it was found that the Ranges of Latitude, Longitude, and Altitude for this particular method was narrower at 0.23, 0.55, and 0.38 meters, respectively, when compared with the method with average measuring only. The tendency of lower data distribution was similar for the case with moving equipment. For the case with non-moving equipment, using more loops to find data averages yielded more stability. As time passed, every increasing Interval of average finding statistically significantly lowered the distribution of data. However, the method of finding averages before implementing Kalman Filter yielded more stable data distribution when N increased compared to implementing Kalman Filter before finding averages. However, statistics showed that when the equipment was entirely still, finding averages and then implementing Kalman Filter at any N, the variances were so close to solely implementing Kalman Filter at any N that the differences were unnoticeable with bare eyes. Similarly, with moving equipment, data measuring for averages before implementing Kalman Filter yielded slightly higher variances compared to the other method; therefore, hardly bearing any effect upon implementation. Nevertheless, the method of implementing Kalman Filter together with measuring for averages with increasing loops was incompatible with the case of moving equipment since the average of a specific position at any time required data from that particular position; otherwise, the results would be incorrect as shown in Picture 7. For example, every loop required 10 meters of a straight line. If data from the current position were combined with data from the previous position 10 meters away and calculated for an average, the result would be the 5-meter average between these two positions, which was 5 meters away from where it was supposed to be. This was the reason why calculations with average loops were unsuitable to be used with moving equipment to lower the variances of GPS positioning data.

TABLE VII. STANDARD DEVIATION AND RANGE OF DATA FROM EACH CALCULATION METHOD

Method	No Movement				Moving	
	GPS Coordinate		True Altitude		True Altitude	
	S.D. (Lat,Lng)	Range (m)	S.D. (m)	Range (m)	S.D. (m)	Range (m)
Raw data	0.0000084, 0.0000159	6.12, 6.78	3.01	16.50	4.66	39.80
KF	0.0000082, 0.0000156	5.67, 6.23	2.97	15.18	4.41	25.44
Find average	0.0000043, 0.0000039	2.34, 2.22	1.04	4.70	2.75	19.50
KF+Average	0.0000042, 0.0000035	2.11, 1.67	1.01	4.32	2.59	17.18
KF+Average every 2 terms	0.0000042, 0.0000035	2.11, 1.67	1.01	4.30	-	-
KF+Average every 3 terms	0.0000041, 0.0000035	2.11, 1.67	1.01	4.28	-	-
KF+Average every 1...663 terms	0.0000012, 0.0000011	0.78, 0.55	0.28	0.97	-	-
Average+KF	0.0000041, 0.0000035	2.11, 1.78	1.01	4.32	2.59	17.18
Average+KF every 2 terms	0.0000040, 0.0000033	1.89, 1.56	0.98	4.08	-	-
Average+KF every 3 terms	0.0000038, 0.0000031	1.78, 1.45	0.95	3.86	-	-
Average+KF every 1...663 terms	0.0000011, 0.0000011	0.78, 0.55	0.29	1.06	-	-

VI. CONCLUSIONS

It can be concluded from this experiment that measuring for averages together with implementing Standard Kalman Filter to three sets of GY-GPS6MV2 Modules to improve the stability of cheap GPS equipment can indeed help reduce the variances of data both when the equipment is constantly moving and when they are completely still. The most effective method is measuring for averages before implementing Standard Kalman Filter. For the case with non-moving equipment, the increasing average loops can lower the variances, whereas, for the case with moving equipment, the increasing average loops reduce data reliability. Even though the increasing loops for average measuring help reduce data variance, it directly varies with time spent collecting data; in other words, the more loops for average measuring, the more time needed for data gathering for return output. From the result, there are limitations of moving measurement. The algorithm will slow down the reding cycle to calculate an average and filter of each reding. That would be the primary direction for future research to overcome these limitations.

In conclusion, this experiment has proved that integrating Standard Kalman Filter with average finding for multiple consumer-grade GPS equipment is another suitable alternative for projects that need to reduce variances from GPS equipment at a lower cost. This innovation can elevate data management with variance through computer commands for technological science and geoinformatics. For example, it can be used with the guidance system searching for missing persons, improving the small projects with customer-grade sensors, or being used to develop future technology and so on continuously.

REFERENCES

- [1] S. Kumar, K.B. Moore, "The evolution of global positioning system (GPS) technology," *Journal of Science Education and Technology*, vol. 11, pp.59–80, March 2002.
- [2] N. Zhu, J. Marais, D. Bétaille, M. Berbineau, "GNSS position integrity in urban environments: A review of literature," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 2761–2778, January 2018.
- [3] J. Park, D. Lee, C. Park, "Implementation of vehicle navigation system using GNSS, INS, odometer and barometer," *Journal of Positioning, Navigation, and Timing*, vol. 4, pp. 141–150, 2015.
- [4] R.J. Meinhold, N.D. Singpurwalla, "Understanding the Kalman filter," *The American Statistician*, vol. 37, No. 2, pp. 123–127, May 1983.
- [5] Z. Li, G. Chang, J. Gao, J. Wang, A. Hernandez, "GPS/UWB/MEMS-IMU tightly coupled navigation with improved robust Kalman filter," *Advances in Space Research*, vol.58, pp. 2424–2434, December 2016.
- [6] K. Feng, J. Li, X. Zhang, X. Zhang, C. Shen, H. Cao, Y. Yang, J. Liu, "An improved strong tracking cubature Kalman filter for GPS/INS integrated navigation systems," *Sensors*, vol.18, no. 6, June 2018.
- [7] B.T.M. Leong, S.M. Low, M.P.L. Ooi, "Low-cost microcontroller-based hover control design of a quadcopter," *Procedia Engineering*, vol. 41, pp. 458 – 464, 2012.
- [8] N. Acosta, J. Toloza, "Techniques to improve the GPS precision," *International Journal of Advanced Computer Science and Applications*, vol.3, no. 8, 2012.
- [9] D. Sathyamoorthy, S. Shafii, Z. Amin, A. Jusoh, S. Ali, "Evaluation of the accuracy of global positioning system (GPS) speed measurement via GPS simulation," *Defence. S&T Technical Bulletin*, vol. 8, no. 2, pp. 121 – 128, November 2015.
- [10] M.G. Wing, "Consumer-grade global positioning systems (GPS) receiver performance," *Journal of Forestry*, vol. 106, pp. 185 – 190, June 2008.
- [11] L. Wang, Z. Li, J. Zhao, K. Zhou, Z. Wang, H. Yuan, "Smart device-supported BDS/GNSS real-time kinematic positioning for sub-meter-level accuracy in urban location-based services," *Sensors*, vol. 16, no. 2, December 2016.
- [12] G.M. Someswar, T. Rao, D.R. Chigurukota, "Global navigation satellite systems and their applications," *International Journal of Software and Web Sciences*, vol. 3, pp. 17 – 23, 2013.
- [13] G. Welch, G. Bishop, *An introduction to the Kalman filter*. University of North Carolina at Chapel Hill, Department of Computer Science, 1995.
- [14] G. Schmitz, T. Alves, R. Henriques, E. Freitas, E. ElYoussef, "A simplified approach to motion estimation in a UAV using two filters," *IFAC-PapersOnLine*, vol. 49, issue. 30, pp. 325 – 330, 2016.
- [15] L. Ljung, "Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems," *IEEE Transactions on Automatic Control*, vol.24, pp. 36 – 50, February 1979.
- [16] K. Reif, S. Gunther, E. Yaz, R. Unbehauen, "Stochastic stability of the discrete-time extended Kalman filter," *IEEE Transactions on Automatic control*, vol. 44, pp. 714 – 728, April 1999.
- [17] A. Benini, A. Mancini, S. Longhi, "An imu/uwb/vision-based extended kalman filter for mini-uav localization in indoor environment using 802.15. 4a wireless sensor network," *Journal of Intelligent & Robotic Systems*, vol. 70, pp. 461 – 476, April 2013.
- [18] D.K. Schrader, B.C. Min, E.T. Matson, J.E. Dietz, "Real-time averaging of position data from multiple GPS receivers," *Measurement*, vol. 90, pp. 329 – 337, August 2016.
- [19] A. Shetty, G.X. Gao, "Measurement Level Integration of Multiple Low-Cost GPS Receivers for UAVs," in *Proceedings of the 2015 International Technical Meeting of the Institute of Navigation*, Dana Point, CA, USA; pp. 26 – 28, January 2015.
- [20] I.K. Ibraheem, S.W. Hadi, "Design and Implementation of a Low-Cost Secure Vehicle Tracking System," in *Proceedings of 2018 International Conference on Engineering Technology and their Applications (ICETA)*, Al-Najaf, Iraq, pp. 146 – 150, September 2018.
- [21] P. Singkhamfu, A. Prasompon "The Accuracy Enhancement of Consumer-Grade Global Positioning System (GPS) for Photogrammetric and City Mapping Determinations," *International Journal of Building, Urban, Interior and Landscape Technology (BUILT)*, vol.14, pp. 81 – 92, 2019.

Artificial Intelligence Model based on Grey Clustering for Integral Analysis of Industrial Hygiene Risk

Alexi Delgado¹, Diana Aliaga², Cristian Carlos³, Lisseth Vergaray⁴, Chiara Carbajal⁵
Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima, Peru¹
Faculty of Environmental Engineering, Universidad Nacional de Ingeniería, Lima, Peru^{2,3,4}
Administration Program, Universidad de Ciencias y Humanidades, Lima, Peru⁵

Abstract—The article proposes a model with an artificial intelligence approach that integrates risks through the Grey Clustering method applying the "Triangulation of center-point based on Whiteness functions -CTWF", for this, the data established is standard data (minimum standards that the four workshops of a company in the industrial sector must meet) and sampled data (real data obtained in the field) to test the grey classes. In this study, the different types of risks (lighting, noise and hand-arm vibration) were globally evaluated and analyzed in the four workshops of a heavy machinery maintenance services company in the industrial sector (welding shop, hydraulic shop, machine shop 1 and machine shop 2), located in Lima, Peru. According to the results obtained from the level of hygienic quality in each workshop, the welding workshop is at a very poor-quality level, while the others are at a good and very good level; regarding the four workshops, it was determined that the noise level is not recommended as they do not meet the minimum required standards. Therefore, control measures were proposed in the four workshops where the level of irrigation is bad and very bad. This study will benefit companies in the industrial sector that need to analyze the level of hygienic quality in their work areas with a global approach in order to apply control measures with prevention, protection of health and physical integrity of workers.

Keywords—Artificial intelligence; grey clustering; industrial hygiene; lighting; noise

I. INTRODUCTION

Actually in the companies of the industrial sector, there can be found diverse types of risks that can affect the zone of comfort and the health of the workers such as excessive level of noise, inadequate illumination, high levels of vibration, etc. [1]; where generally does not exist a method that simultaneously covers these types of risks that will be necessary to obtain indicators of management of safety and occupational health. Therefore the present study proposes a model based on artificial intelligence that integrates these risks through the method of Grey Clustering [2] applied in an industrial sector company located in Lima – Peru, where it is detected that in some of its workshops there are high levels of exposure to hygiene risks such as noise, lighting and vibration; being necessary to have an objective assessment and also propose control measures that is where this study aims.

In this study, risk levels will be evaluated in four workshops of a heavy machinery maintenance service company of the industrial sector (welding workshop, hydraulic workshop, machine workshop 1 and machine workshop 2), located in the province of Lima, Peru, where the results obtained in the occupational hygiene monitoring are observed as high levels of noise, lighting and hand-arm vibration, which will be evaluated globally in each workshop through the Grey Clustering method [3].

In the present study, it is proposed to use the Grey Clustering methodology, which is an artificial intelligence approach [4], to be applied by means of "Centrepont Triangulation based on Whiteness Functions - CTWF" [5], since these are mainly applied to test if the objects of observation belong to predetermined classes, known as grey classes [6], as it is evidenced in the studies of selection of innovative strategies [7], in the evaluation of air quality by grey incidence [8], as well as in the management of occupational safety and health [9]; and as this study will be based on a small group of criteria with limited information, its application will be the most appropriate as grey clustering Method considers this in its analysis.

For this reason, the specific objective of this study is to simultaneously analyze the different types of risks (lighting, noise and hand-arm vibration) present in each workshop, where a global evaluation of the risks will be obtained according to the methodology of Grey Clustering [10] in the heavy machinery maintenance service company of the industrial sector in Peru, and based on the application of the method, control measures can be proposed.

This study is organized as follows: The introduction is presented in Section I. Section II describes the literature review. The methodology is provided in detail in Section III, the case study is described in Section IV; followed by the results and discussions in Section V, and finally in Section VI the conclusions are mentioned.

II. LITERATURE REVIEW

In the study "Evaluation of the noise risk level and its consequences for the technical operators of tobacco processing equipment in a cigarette producing company in Nigeria" the noise generated by machines in three tobacco companies was evaluated and analyzed. Therefore, it was found that all

technical operators are exposed to intense noise above 85 dB, likewise, in the analysis carried out it was shown that all companies operated with efficiencies lower than 55% [11].

In the study "Hearing Loss of Workers Exposed to Noise in a Metalworking Company" the prevalence of hearing loss was determined in 164 workers exposed to noise at levels of 83 to 102 dB, there were analyzed variables such as age, seniority in the job, the use of personal protective equipment and blows to the head, were also included the results of the audiometry made to the workers and the result of the monitoring of noise levels in the work areas. The study determined that 53% of the workers showed normal hearing, while 47% had hearing loss [12].

The objective of the study "Evaluation of lighting levels in interiors and calculation for lighting installations" is to evaluate the level of lighting that the construction workers are exposed to within the workshops of the oil area where the hygienic assessment of physical risk was applied of the Normal Official Mexican. Regarding the lighting conditions in the work centers, the value obtained is compared with what is established in the standards of Executive Decree 2393, where for the calculation of general lighting in interior facilities, the method called General System or the Utilization factor provided average luminance. Likewise, the average evaluation result was 458.22 luxes, which is below the permissible minimum of 500 luxes for design work, revision and correction of plans, so that intervention is necessary through a program to prevent physical risks from lighting [13].

In the study "Response to vibration of the human arm in machine operation", it was carried out in a worker exposed to the vibrations of heavy machines in the industry to determine the risk of musculoskeletal disorders. To examine the effect of these alterations, it was created a 3D model of the human arm in SolidWorks followed by an analysis in ANSYS to obtain the response of the human arm. The analytical and experimental study showed that the degradation of the internal human structure is exponential due to the continuous vibrations that the human body experiences and increases the probability of musculoskeletal disorders [14].

In the study "Evaluation of exposure to vibration risk in the hand-arm segment in companies of the metal-mechanic sector", the accelerations caused by vibrating tools such as: polishers, grinders, drills, among others, were measured and compared with the permissible exposure registered in ISO 5349 of 2002 for the hand-arm-body segment. The study was carried out in four companies of the metal-mechanic sector in the city of Cali. The components studied associated with vibrations were: acceleration, velocity and amplitude. The study showed that the evaluations performed on 15 types of tools did not exceed the permissible limits established by the American Government of Industrial Hygienists of the United States ACGIH 2014, which corresponds to $2.8 m/s^2$ [15].

III. METHODOLOGY

The Grey Clustering method is based on the theory of grey systems. Grey systems study problems with small samples and limited information, where there are currently several problems

of this type [6], this fact makes that grey systems can be applied in different fields, one of them would be in the field of industrial hygiene and safety, since its study in the evaluation of occupational agents usually has this characteristic, in addition this method was applied in different areas such as water management [16], environmental conflicts [17] and the management of occupational safety [9].

The Grey Clustering method was developed to classify observation indices or observation objects into categories using Grey incidence matrices or whitening weighting functions [18]. The method is mainly applied to test whether observation groups belong to predetermined categories.

The present study will use the Grey Clustering method based on central point whitening triangular functions (CTWF) [9]. For that, we have a set of groups "m", and a set of criteria "n" and a set of different grey classes "s", according to the sample x_{ij} ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$) and we have as groups ($i = 1, 2, 3 \dots, m$) and for the criteria ($j = 1, 2, 3 \dots, n$). The CTWF-based grouping of grey classes can be expressed as follows [19], [20]:

Step 1: The individual ranges of the criteria are divided into "s" Grey classes, to then determine the central points of each range in: $\lambda_1, \lambda_2, \dots, \lambda_s$ of Grey classes 1, 2, ..., s of the standard data.

Step 2: Non-dimension of the standard data of each study object (i) and of the selected criteria (j) is carried out. First, the average of the central point's $\lambda_1, \lambda_2, \dots, \lambda_s$ found with step 1 is taken, then each central point will be divided by the mean for each criterion with (1).

$$\lambda_j^k = \frac{\lambda_k}{\frac{1}{s} \sum_{k=1}^s \lambda_k} \quad (1)$$

Step 3: The Grey classes are extended in two directions, adding the Grey classes 0 and (s + 1) with their center points λ_0 and λ_{s+1} respectively. Therefore, the new sequence of central points is set $\lambda_0, \lambda_1, \lambda_2, \dots, \lambda_s, \lambda_{s+1}$ (see Fig. 1). Therefore, the CTWF for the class Grey k^{th} , $k = 1, 2, \dots, s$, of the criterion j^{th} , $j = 1, 2, \dots, n$, for an observed value x_{ij} is defined by (2).

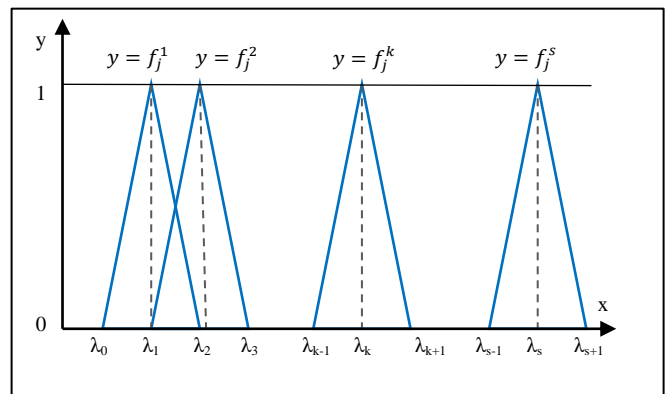


Fig. 1. Center-Point based on Whitening Functions – CTWF [20].

$$f_j^k = \begin{cases} 0; x \notin [\lambda_0, \lambda_1] \\ \frac{x-\lambda_{k-1}}{\lambda_k-\lambda_{k-1}}; x \in [\lambda_{k-1}, \lambda_k] \\ \frac{\lambda_{k-1}-x}{\lambda_{k+1}-\lambda_k}; x \in [\lambda_k, \lambda_{k+1}] \end{cases} \quad (2)$$

Where: is the CFW of the k^{th} Grey class of the j^{th} criterion and is the weight of the j criterion.

Step 4: The weights of the criteria are calculated using the harmonic mean, first the inverse of the dimensionless or non-dimension standard data must be determined and divided by the sum of the inverses found, it will be calculated using (3).

$$n_j^k = \frac{\frac{1}{\lambda_j^k}}{\sum_{j=1}^m \frac{1}{\lambda_j^k}} \quad (3)$$

Step 5: The Clustering coefficient σ_i^k , which indicates the weight of the criteria, for group $i, i = 1, 2, \dots, m$, with respect to the Grey class $k, k = 1, 2, \dots, s$ is calculated by (4).

$$\sigma_i^k = \sum_{j=1}^n f_j^k(x_{ij}) \cdot n_j \quad (4)$$

Step 6: If $\max_{1 \leq k \leq s} \{\sigma_i^k\} = \sigma_i^{k^*}$, we decide that the object belongs to the class Grey k^* . However, when there are several objects in the Grey class k^* , these objects can be ordered according to the magnitudes of their Clustering coefficients and we stretch the result for each object of study.

IV. CASE STUDY

The application of the method will be carried out in 4 areas of a company of maintenance services of heavy machinery, considering three parameters (Noise, Lighting and vibration) for the analysis of the conditions of industrial hygiene, the company is located in the industrial avenue, Cercado de Lima – Peru.

Fig. 2 shows the location of Peru and Lima region where the study was conducted.

In Fig. 3, the locations of the four study areas within the industrial plant are shown.



Fig. 2. Location of Peru and the Lima Region.

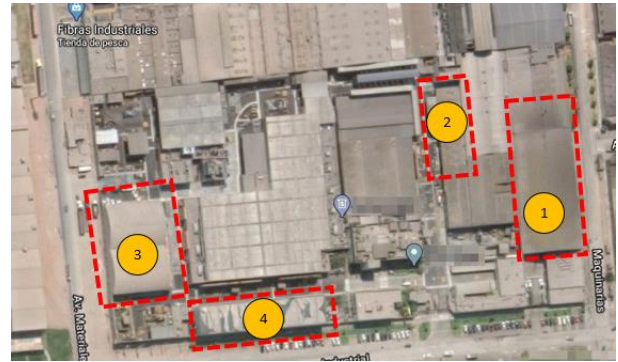


Fig. 3. Location of the Workshops to be Analyzed: (1) Welding Workshop, (2) Hydraulic Workshop, (3) Machine Workshop 1 and (4) Machine Workshop 2. Source: Google Maps.

A. Definition of Study Objects

The study will be based on the data obtained from the monitoring reports of occupational hygiene in four workshops of a heavy machinery maintenance service company. In Table I, the objects are defined by assigning them an abbreviated notation.

TABLE I. STUDY OBJECTS' NOTATIONS

Objects	Notation
Welding workshop	G1
Hydraulic Workshop	G2
Machine Workshop 1	G3
Machine Workshop 2	G4

B. Definition of Evaluation Criteria

The criteria for this study are defined in Table II.

In Table III, the average data of the criteria obtained from the reports of occupational hygiene realized in the month of September and October 2019 is presented, whose values are projected for a working day of 8 hours.

C. Definition of Grey Classes

In Table IV, the grey classes that have been defined in four ranges and will represent the levels of hygienic quality in the different objects of studies mentioned in the Table I.

TABLE II. CRITERIA WITH THEIR RESPECTIVE UNIT AND NOTATION

Criteria	Unit	Notation
Noise	Decibel	C1
Illumination	Lux	C2
Vibration	m/s2	C3

TABLE III. AVERAGE VALUES OF THE CRITERIA OBTAINED FROM THE OCCUPATIONAL HYGIENE REPORTS OF THE WORKSHOPS

Objects	C1 (dB)	C2 (Lux)	C3 (m/s2)
G1	98.18	615.20	4.196
G2	91.18	463.73	1.783
G3	96.40	649.72	1.750
G4	97.40	1131.75	1.085

TABLE IV. GREY CLASSES' NOTATIONS

Level	Notation
Very poor	λ_1
Poor	λ_2
Good	λ_3
Very good	λ_4

Also, the ranges of the criteria are divided in 4 Grey classes $\lambda_1, \lambda_2, \lambda_3$ and λ_4 , are shown in Table V. To establish the intervals of the grey classes for each criterion, used national standards such as RM No. 375 - 2008 -TR [21], and documents from international organizations such as the Encyclopedia of Occupational Safety and Health in the work of OIT [22].

TABLE V. RANKS OF THE CRITERIA AND THEIR 4 GREY CLASSES

Criteria	Grey Classes			
	Very poor	Poor	Good	Very good
C1: Noise	85 - 100	70 - 85	65 - 70	60 - 65
C2: Illumination	100 - 300	300 -500	500 - 800	800 - 1000
C3: Vibration	4 - 6	2.5 - 4	1 - 2.5	0 - 1

D. Calculations with Grey Clustering

Step 1: Determine the center points. The central points are obtained from the semi-sum of the extremes of each range shown in Table V, and these values are shown in Table VI.

TABLE VI. CENTRAL POINTS OF THE CRITERIA FOR EACH GREY CLASS

Criteria	Grey Classes			
	Very poor	Poor	Good	Very good
C1: Noise	92.5	77.5	67.5	62.5
C2: Illumination	200	400	650	900
C3: Vibration	5	3.25	1.75	0.5

Step 2: Non-dimension. Table VII shows the results obtained from the non-dimensioning of the standard data by means of (1).

TABLE VII. DIMENSIONLESS VALUES OF STANDARD DATA

Criteria	$f_j^1(x)$	$f_j^2(x)$	$f_j^3(x)$	$f_j^4(x)$
C1: Noise	1.23	1.03	0.90	0.83
C2: Illumination	0.37	0.74	1.21	1.67
C3: Vibration	1.90	1.24	0.67	0.19

Table VIII shows the results obtained from the sizing of the sampling data, which is obtained by dividing each data by the mean of the standard data for each criterion.

TABLE VIII. DIMENSIONLESS VALUES OF THE SAMPLING DATA

Criteria	G1	G2	G3	G4
C1: Noise	1.31	1.22	1.29	1.30
C2: Illumination	1.14	0.86	1.21	2.11
C3: Vibration	1.60	0.68	0.67	0.41

Step 3: Determination of triangular functions and values. The whitenization functions are elaborated for each evaluation criterion with the data shown in Table VII, which is the reason three functions are considered (noise, illumination, and vibration). Next, the function for the noise criterion is presented (see Fig. 4).

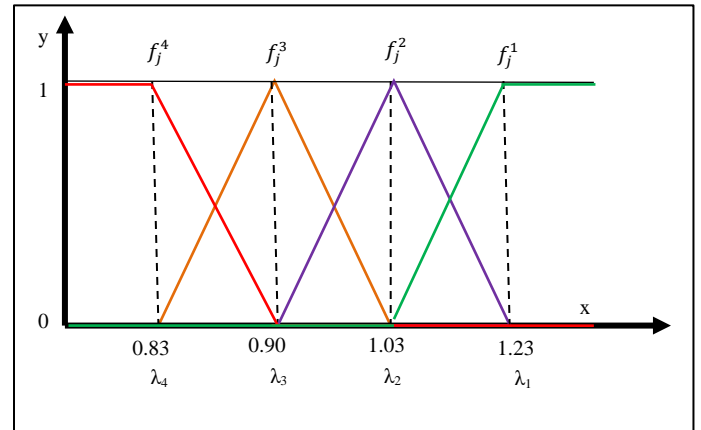


Fig. 4. Whitenization Functions for Criterion 1.

As an example, the functions corresponding to criterion 1 (Noise) are presented in (5) – (8).

$$f_j^1 = \begin{cases} 1; & x \in [1.23; +\infty) \\ \frac{1.03-x}{1.03-1.23}; & x \in (1.03; 1.23) \\ 0; & x \in [0; 1.03] \end{cases} \quad (5)$$

$$f_j^2 = \begin{cases} 0; & x \in [0; 0.90] \cup [1.23; +\infty) \\ \frac{x-0.90}{1.03-0.90}; & x \in (0.90; 1.03] \\ \frac{1.23-x}{1.23-1.03}; & x \in (1.03; 1.23) \end{cases} \quad (6)$$

$$f_j^3 = \begin{cases} 0; & x \in [0; 0.83] \cup [1.03; +\infty) \\ \frac{x-0.83}{0.90-0.83}; & x \in (0.83; 0.90) \\ \frac{1.03-x}{1.03-0.90}; & x \in (0.90; 1.03) \end{cases} \quad (7)$$

$$f_j^4 = \begin{cases} 1; & x \in [0; 0.83) \\ \frac{0.90-x}{0.90-0.83}; & x \in (0.83; 0.90) \\ 0; & x \in [0.90; +\infty) \end{cases} \quad (8)$$

The functions of the other criteria are developed following the same procedure as above.

In Table IX, the results shown in the welding workshop, after evaluating non-dimensioning sampling values (Table VIII) in the functions of each criterion.

TABLE IX. VALUES OF THE FUNCTIONS EVALUATED IN THE WELDING WORKSHOP

Criteria	C1	C2	C3
f_j^1	1.00	0.00	0.55
f_j^2	0.00	0.15	0.45
f_j^3	0.00	0.85	0.00
f_j^4	0.00	0.00	0.00

Table X shows the results of the hydraulic workshop, after evaluating the non-dimensioning values of Table VIII in the functions of each criterion.

TABLE X. VALUES OF THE FUNCTIONS EVALUATED IN THE HYDRAULIC WORKSHOP

Criteria	C1	C2	C3
f_j^1	0.95	0.00	0.00
f_j^2	0.05	0.74	0.018
f_j^3	0.00	0.26	0.982
f_j^4	0.00	0.00	0.00

In Table XI, the results shown in the machine shop 1 after evaluating non-dimensioning sampling values of Table VIII in the functions of each criterion.

TABLE XI. VALUES OF THE FUNCTIONS EVALUATED IN MACHINE WORKSHOP 1

Criteria	C1	C2	C3
f_j^1	1.00	0.00	0.00
f_j^2	0.00	0.00	0.00
f_j^3	0.00	1.00	1.00
f_j^4	0.00	0.00	0.00

In Table XII, the results shown in the machine shop 2 after evaluating the non-dimensioning sampling values (Table VIII) in the functions of each criterion.

TABLE XII. VALUES OF THE FUNCTIONS EVALUATED IN MACHINE WORKSHOP 1

Criteria	C1	C2	C3
f_j^1	1.00	0.00	0.00
f_j^2	0.00	0.00	0.00
f_j^3	0.00	0.00	0.46
f_j^4	0.00	1.00	0.54

Step 4: We determine the weight of the criteria. In Table XIII, the weights of the criteria are shown; for this, (3) is used and the calculations are made for each evaluated workshop.

TABLE XIII. CRITERIA WEIGHTS FOR EACH WORKSHOP EVALUATED

Weights	$f_j^1(x)$	$f_j^2(x)$	$f_j^3(x)$	$f_j^4(x)$
C1	0.20	0.31	0.32	0.17
C2	0.67	0.43	0.24	0.08
C3	0.13	0.26	0.44	0.74

Step 5: Determine the clustering coefficient. In Table XIV shows the results of the clustering coefficients, obtained through the application of (4). The shaded boxes indicate the maximum clustering value for each workshop, which indicates the level of hygienic quality it presents.

Step 6: Results using Max clustering coefficient. Table XV shows a summary with the maximum clustering coefficients

that indicates the level of hygienic quality with respect to the three criteria established in the workshops.

TABLE XIV. CLUSTERING COEFFICIENT FOR EACH WORKSHOP

	Very poor	Poor	Good	Very good
	$f_j^1(x)$	$f_j^2(x)$	$f_j^3(x)$	$f_j^4(x)$
G1: Welding Workshop	0.27	0.18	0.20	0.00
G2: Hydraulic Workshop	0.19	0.34	0.49	0.00
G3: Machine Workshop 1	0.20	0.00	0.68	0.00
G4: Machine Workshop 2	0.20	0.00	0.20	0.49

TABLE XV. MAXIMUM CLUSTERING COEFFICIENT FOR EACH WORKSHOP

Group	σ_i^k	Hygienic Quality Level
G1: Welding Workshop	0.27	Very poor
G2: Hydraulic Workshop	0.49	Poor
G3: Machine Workshop 1	0.68	Good
G4: Machine Workshop 2	0.49	Very good

V. RESULTS AND DISCUSSION

A. About the Case Study

According to the results obtained in Table XV, it can be observed that the quality level of hygiene of machine shop 2 is very good; likewise, machine shop 1 and hydraulic shop present a good level of quality and the welding workshop presents a very bad level of hygienic quality with respect to the established parameters.

In Table XVI, the analysis of each workshop is made with its respective criteria (noise, illumination, and vibration). It can be observed that in the 4 workshops with respect to the noise level they are not at the recommended level, that is, they do not comply with the minimum standards according to RM 375-2008 TR.

In the present paper, the Grey Clustering methodology was used, which showed that the four workshops analyzed have a very poor level of noise quality, giving us an overview to visualize the risk in the areas. On the other side, in the study realized by Xingsong, W et al [11], noise measurements were taken and questionnaires were conducted to assess the psychological impact and effects of noise on workers, which showed that workers are exposed to noise in excess of 85 dB.

TABLE XVI. RESULTS OF EACH WORKSHOP WITH THEIR RESPECTIVE CRITERIA

Group	Noise	Illumination	Vibration
G1: Welding Workshop	Very poor	Good	Poor
G2: Hydraulic Workshop	Very poor	Poor	Good
G3: Machine Workshop 1	Very poor	Good	Very Good
G4: Machine Workshop 2	Very poor	Very Good	Very Good

In the study conducted by Machado, M et al [13], a methodology was applied to evaluate the risk of lighting in the workshops of an oil area where there was evidence of low lighting levels, as in this study, the hydraulic workshop presented the same situation. The methodology of the mentioned study mentioned with the methodology applied in our study is different in that only the physical agent of illumination can be evaluated, it does not apply for more criteria of integral risk evaluation as in the case of Grey Clustering.

In the study [15], vibration was measured for vibrating tools of 4 companies in the metal-mechanical sector, it was found that 15 of them did not exceed the permissible limits. For this study four workshops of a heavy machinery company were analyzed, finding the quality levels of each one of them, in which a workshop with a bad quality level was found; instead, the mentioned study focused only on those tools that presented a high vibration level for each company, which would have been better to apply the Grey Clustering method to obtain an integral vibration risk in each one of them.

B. About the Methodology

The Grey Clustering methodology helped us to determine the deficient points of the criteria present in each Workshop as well as the level of quality of comfort in these workshops. It is observed in Table XVII the advantages and disadvantages in the Delphi, AHP methods and FAHP [23].

TABLE XVII. ADVANTAGES AND DISADVANTAGES OF EXISTING METHODS FOR APPLICATION IN DIFFERENT STUDIES

Method	Advantage	Disadvantages
Grey clustering	<ul style="list-style-type: none"> Evaluate problems with small groups of criteria. Objective Analyze with limited information and uncertainty. 	<ul style="list-style-type: none"> Does not cover complex problems.
DELPHI	<ul style="list-style-type: none"> Expert group estimates based on experience. Application in any field whether political, legal, educational 	<ul style="list-style-type: none"> It is subjective Appearance of bias in the participants (specialists) Intuitive
AHP process analytical hierarchy	<ul style="list-style-type: none"> Evaluate complex problems with multicriteria. Objective 	<ul style="list-style-type: none"> It requires the participation of the knowledge and experiences of the experts who will make decisions about the problem
FAHP	<ul style="list-style-type: none"> Evaluate various problems with multicriteria Objective 	<ul style="list-style-type: none"> It must have precise and well-defined information, that is, it is based on fuzzy logic.

C. Control Proposals

1) *Group G1:* Welding workshop: Noise (Very poor quality level): As the main measure, it is proposed to realize an engineering control based on wave interference by installing equipment that eliminates the sound pressure level by more than 30 decibels, while the project is being developed it is recommended to provide the workers with personal protection equipment. Vibration (Poor quality level): It is

proposed to implement anti-vibration handles for manual tools, in addition to reduce the exposure time of the workers through staff rotation and providing anti-vibration handles (as shown in Fig. 5).

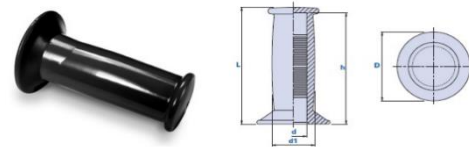


Fig. 5. Anti-Vibration Handle.

2) *Group G2:* Hydraulic Workshop: Noise (Very poor-quality level): It is proposed to offer double hearing protection (earmuff + earplug) to workers in order to reduce exposure to noise in the work environment. Illumination (Bad Quality Level): It is proposed as an engineering control measure to redesign the illumination system to reduce the risk level of illumination and to comply with the minimum standards required in the workshop (see Fig. 6).

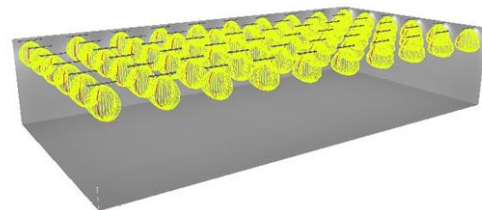


Fig. 6. Luminaire Distribution Reference.

3) *Group G3:* Machine workshop 1: Noise (Very poor quality level): It is proposed to provide double hearing protection (earmuff + earplug) according to the model shown in Fig. 7, to the workers in order to reduce exposure to noise in the work environment.



Fig. 7. Model of Recommended Earmuffs (PELTOR X4P5E- 3M) and Earplugs (3M E-A-Rsoft Yellow Neons).

4) *Group G4:* Machine workshop 2: Noise (Very poor-quality level): It is proposed to provide double hearing protection (earmuff + earplug) according to the model shown in Fig. 7 to the workers to reduce exposure to noise in the work environment.

VI. CONCLUSIONS

In the present work it was found that, of the 4 industrial workshops analyzed, Machine Shop 2 represents a Very Good hygienic quality level, Hydraulics and Machine Shop 1 a Good quality level, while the welding workshop represents a Very Poor-quality level. It was also observed that, in the Hydraulics

workshop, the lighting criterion represents a Poor-quality level, in the welding workshop the vibration criterion represents a Poor-quality level, and finally for the noise criterion all areas represent a Very Poor quality level. The results obtained made it possible to propose control measures aimed at improving working conditions in the company.

The grey clustering methodology was applied, which has proved to be effective because it allowed obtaining the results of the hygienic risk in a global way and also in a specific way of the areas for each criterion that was established. This allows us to improve the management of occupational hygiene in the companies allowing to establish control measures before the risk found to reduce it to a level of better hygienic quality and improve the health of the workers, for this study this method is more effective compared to the DELPHI, AHP and FAHP methods; therefore, it is effective for the risk analysis defining groups and criteria.

In future research, it is recommended to apply the Grey Clustering method to obtain a global risk of hygienic quality in sectors such as mining, hydrocarbons, construction, and manufacturing, among others, thus continuing to improve working conditions and obtain healthy working environments. This method could also be applied to verify that the Industrial Hygiene and Safety programs are being developed effectively, otherwise further improvements will need to be implemented.

REFERENCES

- [1] F. H. Robert, "Higiene industrial," *Encicl. Salud y Segur. en el Trab.*, p. 38, 2000.
- [2] D. Julong and Y. Lin, "Introduction to grey systems theory," *Underst. Complex Syst.*, vol. 68, pp. 1–24, 1988, doi: 10.1007/978-3-642-16158-2_1.
- [3] A. Delgado, P. Montellanos, and J. Llave, "Air quality level assessment in Lima city using the grey clustering method," *Jan.* 2019, doi: 10.1109/ICA-ACCA.2018.8609699.
- [4] A. Delgado and I. Romero, "Environmental conflict analysis on a hydrocarbon exploration project using the Shannon entropy," in *Proceedings of the 2017 Electronic Congress, E-CON UNI 2017, 2018*, vol. 2018-Janua, doi: 10.1109/ECON.2017.8247309.
- [5] A. Delgado, J. Culqui, G. Tasayco, A. Millán, E. Tirado, and C. Carbajal, "Quality assessment of surface water associated with a copper mine in peru using grey systems," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 6660–6668, 2020, doi: 10.30534/ijatcse/2020/359942020.
- [6] L. N. Zhang, F. P. Wu, and P. Jia, "Grey Evaluation Model Based on Reformative Triangular Whitenization Weight Function and Its Application in Water Rights Allocation System," *Open Cybern. Syst. J.*, vol. 7, no. 1, pp. 1–10, 2013.
- [7] Ling P. Ling and xin W. Zheng, "An optimized grey cluster model for evaluating quality of labor force," *J. Softw.*, vol. 8, no. 10, pp. 2489–2494, 2013, doi: 10.4304/jsw.8.10.2489-2494.
- [8] A. Delgado, E. Luna, M. Hernández, K. Montero, and C. Carbajal, "Assessment of the air quality in four cities with near mining activity in mexico, using the grey clustering method," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, pp. 7514–7518, Sep. 2019, doi: 10.35940/ijrte.C5696.098319.
- [9] A. Delgado, J. Maguiña, R. Cabezas, S. Hidalgo, and C. Carbajal, "Integral assessment of risk level in libraries using the grey clustering method," *Int. J. Recent Technol. Eng.*, vol. 8, no. 3, 2019, doi: 10.35940/ijrte.C5695.098319.
- [10] A. Delgado and I. Romero, "Social impact assessment on a hydrocarbon project using triangular whitenization weight functions," 2016, doi: 10.1109/CACIDI.2016.7785998.
- [11] X. Wang, O. A. Orelaja, D. S. Ibrahim, and S. M. Ogbonna, "Evaluation of noise risk level and its consequences on technical operators of tobacco processing equipment in a cigarette producing company in Nigeria," *Sci. African*, vol. 8, p. e00344, 2020, doi: 10.1016/j.sciaf.2020.e00344.
- [12] B. Z. González, V. P. Sierra, J. I. V. Martínez, Y. C. Muraira, and V. R. Catalina, "Disminución Auditiva de Trabajadores Expuestos a Ruido en una empresa Metalmeccánica," *Ciencia Trab.*, vol. 35, no. April 2016, pp. 233–236, 2010.
- [13] E. T. Machado Miranda, S. E. Nuela Sevilla, A. P. López-López, and D. L. Mosquera Guanoluusa, "Evaluación niveles de iluminación en interiores y cálculo para instalaciones de alumbrado," *KnE Eng.*, vol. 2020, pp. 13–36, 2020, doi: 10.18502/keg.v5i2.6215.
- [14] M. E. Abdullah, M. A. Hassan, and A. Israr, "Vibration response of human arm in machine operation," 6th Int. Conf. Aerosp. Sci. Eng. ICASE 2019, 2019, doi: 10.1109/ICASE48783.2019.9059121.
- [15] C. Arias, J., Martínez, "Evaluación de la exposición al riesgo por vibraciones en el segmento mano brazo en compañías del sector metalmeccánico," *Med. Segur. Trab. (Madr.)*, vol. 62, no. 245, pp. 327–336, 2016.
- [16] A. Delgado, A. Aguirre, E. Palomino, and G. Salazar, "Applying triangular whitenization weight functions to assess water quality of main affluents of Rimac river," *Proc. 2017 Electron. Congr. E-CON UNI 2017*, vol. 2018-Janua, pp. 1–4, 2018, doi: 10.1109/ECON.2017.8247308.
- [17] V. Bax, W. Francesconi, and A. Delgado, "Land-use conflicts between biodiversity conservation and extractive industries in the Peruvian Andes," *J. Environ. Manage.*, vol. 232, pp. 1028–1036, Feb. 2019, doi: 10.1016/j.jenvman.2018.12.016.
- [18] L. Sifeng and L. Yi, *Grey Systems, Theory and Applications*. Chennai, India: Springer, 2010.
- [19] S. Liu and Y. Lin, "Grey Information: Theory and Practical Applications," New York, 2006.
- [20] S. Liu and Y. Lin, *Grey Systems: Theory and Applications*. Berlin: Springer, 2010.
- [21] R. M. N. 375-2008-TR, "Norma Básica de Ergonomía y de Procedimiento de Evaluación de Riesgo Disergonómico," pp. 1–17, 2008.
- [22] M. J. Griffin, "Vibraciones," *Encicl. Salud Y Segur. En El Trab. en la OIT*, vol. II, p. 18, 2012.
- [23] A. Delgado and H. Flor, "Selection of the best air purifier system to urban houses using AHP," in *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, 2017*, vol. 2017-Janua, doi: 10.1109/DISTRA.2017.8229622.

Intelligent Traffic Light Controller using Fuzzy Logic and Image Processing

Abdelkader Chabchoub¹, Ali Hamouda², Saleh Al-Ahmadi³, Adnen Cherif⁴

Physics Department, Faculty of Sciences of Tunis - University of Tunis El Manar, Tunis, Tunisia^{1,4}

Electronics Department, Madinah Technical College and Interserve Learning & Employment, Leeds, United Kingdom²

Electrical Department, Faculty of Engineering, Islamic University of Medinah, Medinah, KSA³

Abstract—Today's traffic congestion in the big city is a serious problem, as it causes a lot of environmental pollution and difficulty in transportation, which leads to difficult daily life for the human beings in addition to material losses. In this work a smart traffic light controller was designed using fuzzy logic and image processing with MATLAB, to control movement in two ways, aided by a camera and auto sensors. The Fuzzy logic has two inputs and six outputs designed, the console input is the number of cars on each road and the time of the assumed red, yellow and green signal according to the vehicles congestion. The simulation result is similar to the proposed control unit, as it deals with the lights simultaneously according to the number of cars in each branch of the road, which leads to the use of all the time to operate the stoplights. Our system can be employed in solving the problem of traffic congestion in the big cities or the smart cities.

Keywords—Traffic congestion; smart city; traffic light; fuzzy logic; image processing; objects detections

I. INTRODUCTION

Traffic congestion in many modern cities around the world is a severe problem. Many critical problems and challenges are caused by traffic congestion in the large and densely populated urban areas moving to different places with the abundance of cars in crowded cities will become more difficult [11]. Due to these traffic problems, people lose time, miss opportunities and get disappointed. Overcrowding directly impacts companies. Due to that, there is a loss in the productivity of workers, and opportunities are lost, delivery gets delayed and thereby the cost goes on increasing. To solve efficiently these congestion problems, we designed an intelligent traffic control system. The essential goals of this paper are to improve safety, minimize travel time and increase the capacity of infrastructures [6]. Our case in Al Madinah, Saudi Arabia is that there are unlimited visitors during the Haj and Umra seasons, which increases the traffic congestion. A proposed fuzzy controller to control the car stream and traffic congestion [10] is related to the number of cars and population and road size in Al Madinah and Makah cites. They are a different case. First at the end of working days, in the morning, during the end of the weeks and holidays, a huge number of people come to the Al Haram. In addition, the unlimited number of cars which cause traffic congestion, and the number of visitors will increase in Ramadan and Haj seasons. For these reasons, traffic control is a big challenge for these two cites. Everyone can repeatedly see that there are really traffic problems. Our proposal is to solve this problem via using artificial intelligent.

Our controller is designed with fuzzy logic control [9], and at each stop there are multiple sensors and cams to capture photos and images. The processing algorithm is used to detect the number of vehicles in each direction. And we used an optical sensor equipped with cameras to detect the number of cars coming to the traffic light, and another to detect the number of cars leaving. The number of cars in the signal can be calculated by subtracting the number of cars entering from the number of the existing ones [12]. This process is repeated for each road, and the total entries for the first and second roads are the fuzzy logic inputs. The fuzzy Controller is designed to estimate the traffic time according to the number of cars at each road irrespective of the fixed time [1]. The simulation result shows an excellent result, and the program can be designed and run-on microcontroller of PLC and/or any other controller. Also, instead of an optical sensor, a magmatic or ultrasonic sensor can be used. Many researchers and papers focused on the study of traffic light and control. Conventional traffic light control system currently handles traffic at one junction. However, the synchronization of traffic light systems still caused congestion within them. This work proposes a system based on a microcontroller that controls the intensity of traffic using infrared sensors and achieves dynamic time slots at different levels. Also, the portable controller will solve the problem of cars stuck in traffic, which will benefit the economy, society and the environment. This paper discusses a solution that uses a Special Purpose Simulation Tool (SPS) to improve signal light timing at multiple signal and intersections [2]. Autonomous vehicle and driver assistance systems use various sensors such as sensor and radar to detect their surroundings, but they cannot detect standard traffic lights. To solve this problem, a previous map booth is used to predict the location of traffic lights.

Traffic congestion due to inadequate space and funds has led researchers to think of a solution to reduce it. The solution is using intelligent system. One of them is the use smart traffic light (STL) and wireless sensors network (WSN) [3]. The WSN collects data about traffic lanes in real time in terms of traffic quantity (TQ) and waiting time (WT). It then computes a priority degree (PD) that determines order of green light assignment [4]. For the past years, traffic signal control system was static and not efficient. For a better traffic control, there should be a more efficient and dynamic system that handles traffic easily and more safely. This system will be better particularly in the performance of traffic intersections control [5]. The main objective for this study is to introduce a new traffic signal controller based on fuzzy logic. In the second part

of the study, the before-after measurements are introduced. This paper proposes a traffic signal synchronization system which inputs real-time data using fuzzy logic. The paper also uses Q-learning module, so the system learns by itself by updating the set of role base [8].

This paper is organized as follows. In Section 2, the proposed modules will be presented with, in particular the intelligent fuzzy logic controller design. Section 3 describes the membership function Triangular types of function are used. Section 4 presents the fuzzy rule base and establishment also the result of MATLAB simulation.

II. PROPOSED MODULES

The present traffic light depends on fixed time for each road. The first road, Ro1, has three signals (R1, G1 and Y1); the second road has three signals (R2, G2 and Y2). The controller is programmed with fixed time for each pair (R1 and G2) having the same time. Also (R2 and G1), the third signals yellow once both (Y1 and Y2) have the same timing at all time. The duration of each one is with a fixed time all the time, and it will be monitored just manually by a policeman. Our proposed system (artificial intelligent control) is fuzzy logic controller with multiple sensors and cams distributed around the traffic signal at each direction. The function of each sensor is to sense if there is a car or not and if any car entered the target zone, in which case the sensor activates the camera to capture a photo, and the photo is processed and the number of cars on the photo are calculated. The same mechanism is applied to each road, and the signal time for the road is not fixed. It depends on the number of cars on each road at a time, and the time must be set by the fuzzy logic controller which depends on the design of the fuzzy rules and the number of cars at each road. Fuzzy logic has been used extensively to develop a traffic light controller because it allows the qualitative modeling of complex systems that are difficult to solve using classic mathematical models. It is also good for systems that have multiple changes at a specific time. Several researchers have proposed traffic light control systems using fuzzy logic. They proposed FLSC for a 4-way isolated intersection from East / West / North and South without moving traffic. This generally provides better performance for FLSC compared to fixed time and actuated controllers. Today, all FLSC research work has developed based on unstable traffic conditions especially in developed countries, [7]. The main objective of this research is to design an intelligent fuzzy logic controller to control four way traffic lights. MATLAB fuzzy logic and image processing tools books were used for simulations to examine and analyze the effectiveness of the proposed FLSC. Then, the optimal performance of the proposed controller is typically contrasted with an optimized fixed time.

A. Design Criteria and Constraints

The following assumptions have been made to develop a fuzzy logic control system for traffic lights:

- Interchange: It is a four-way intersection between the traffic coming from the direction of the first road (north, south) and the second road (west and east).
- When traffic moves from north and south it stops from west and east and vice versa.

- The left and right turns are not considered.
- Fog logic unit inputs from the first and second paths taken from sensors and cameras.
- The picture is taken from every point of view and is usually from several cameras.

B. Design of Fuzzy Logic Controller

Two fuzzy logic controllers were designed for controlling the flow of two ways (road one RO1 and Road two RO2) Fig. 1, with six traffic signals (green one, green two, red one, red two, yellow one and yellow two). The number of cars on road one can be calculated by fixing a sensor 30m from the signal. Road one is assumed with three sub roads with full capacity of 24 cars, and road two is assumed with two sub roads with maximum capacity of 16 cars as represented of fuzzy logic input. There are many sensors at each road, and the function of the sensor is to activate the camera to capture cars photos then process them to calculate the actual number of cars in the specific zone and activate the controller to adjust signals time.

C. Fuzzy Logic Input and Image Processing

The number of cars is a very important and complicated topic. If any car appears in the cams zoom, the first optical sensor or ultrasonic or metal detection sensor is used to activate the camera to capture a photo. For the specific, zoom in which the car is detected, then it is sent to the processor to filter, and is converted to the black and white color. Then the surrounded edge is detected as in Fig. 2 and the inter space is filled with identified color to calculate the number of cars. We can calculate the number of objects on the processed image [13].

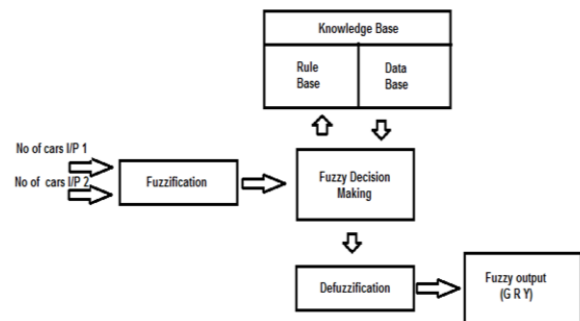


Fig. 1. Fuzzy Logic Controller.

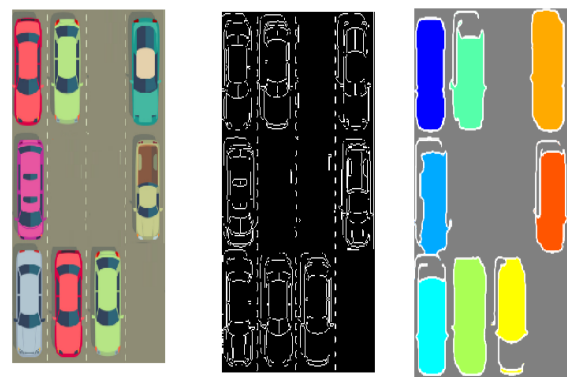


Fig. 2. Image of Cars on the Roads and Steps of Image Processing Filtering and Objects Detection.

III. MEMBERSHIP FUNCTION

In this work, we used a triangular membership function due to their computational efficiency. The membership function for each input-output ambiguous variable in FLSC is as follows:

- Membership function for the number of cars on two roads.
(Road one RO1 and Road two RO2)
- Membership function for the output traffic signal.
(R1, Y1, G1, R2, Y2 and G2)

The number of cars on the first input Road one RO1 is {C1, C2, C3, C4, C5, C6, C7 and C8}, the second roads RO2 is {C11, C22, C33, C44, C55}, the total capacity of the cars on the roads covered by the cams zooms are assumed to be 24 on RO and 15 on RO2, respectively. Traffic signal time has linguistic variables as in the Fig. 3. The potential membership function of the number of vehicles standing in line is represented in the line at traffic lights as shown in Fig. 4. After the sensor activates the camera to capture photos from the road, it is processed in many steps: first converted to black and White photo, filtered and edged, detected and filed as objects then functions are used to calculate objects on the road as in Fig. 2.

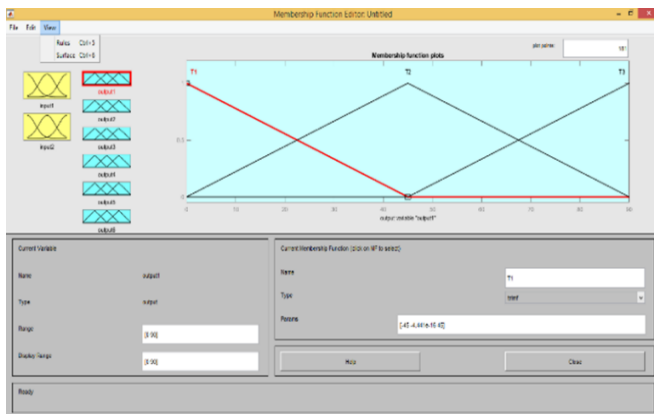


Fig. 3. Membership Function the Time of First Output Signal Divided into Three Time Zone (T1, T2 and T3).

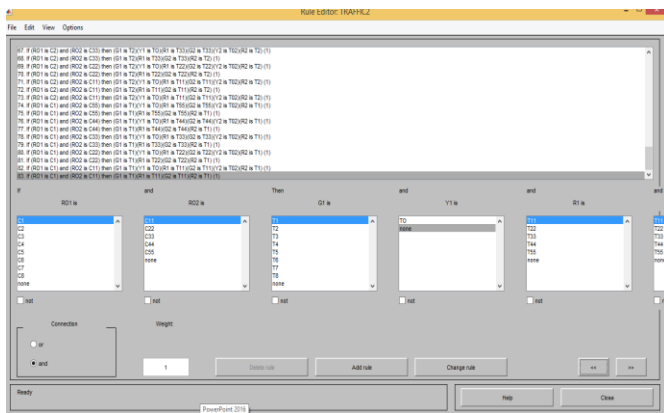


Fig. 4. Rules Viewer for the Fuzzy Logic with Two Inputs (Cars on RO1 and RO2) and Six Output are the Timing of the Traffic Signal (R1, G1, Y1, R2, G2 and Y2). Each Time of the Signal can be Determined by Fuzzy Logic Controller and its Function of the Number of Cars at each Roads.

IV. FUZZY RULE BASE AND ESTABLISHMENT

The basic function of the fuzzy rule base is to represent the expert knowledge in a form of IF-THEN rule structure combined with AND/OR operators [8]. For e.g., IF the number of cars in road one is C3 and in the second road is C11, then the output of traffic signal is adjusted according to the number of cars in each roads Fig. 5. The fuzzy rule base is set of fuzzy rules. It maps the combination of fuzzy inputs (number of cars on each roads) to the corresponding fuzzy output (signal times of RYG) Fig. 6. In this paper we consider different membership of cars number and different time zoon, changed according to the Traffic congestion.

In our work this with different inputs, the extension time (z axis) is small when the access density (y axis) is small and the queue side density (x axis) is also small Fig. 7. This indicates profit for time, unlike the other method, where on the three axes the growth is external time. In other words, the outside time grows slowly because it is large only when the reach side density increases, and the intensity of the queue side is constant. Another advantage is if the access side density is constant and the queue side density increases, then the rollover time shifts to medium to short. This is an important difference between the other methods, especially in the fixed time system.

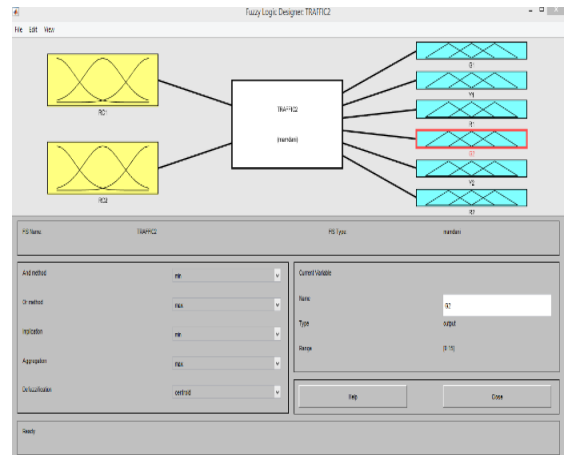


Fig. 5. Fuzzy Logic Controller with Two Inputs Cars on Two Roads and Six Output are the Timing of the Traffic Signal (R1, G1, Y1, R2, G2 and Y2).

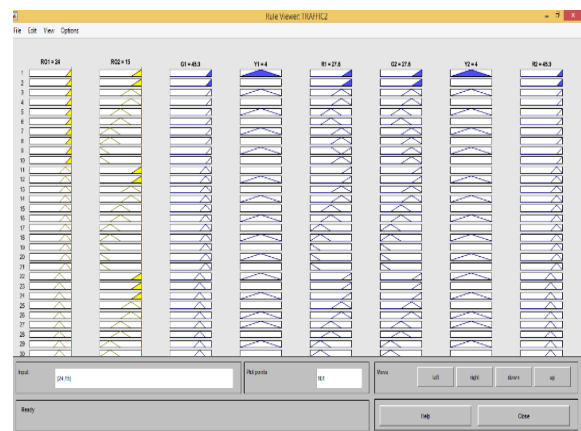


Fig. 6. Rules Viewer Representing the Input and Output of each Roads, if they are 24 Cars on RO1 and 15 Cars on Road Two then $G1=R2=45.9$ sec, $Y1=Y2=4$ sec and $G2=R1=27.6$ sec.

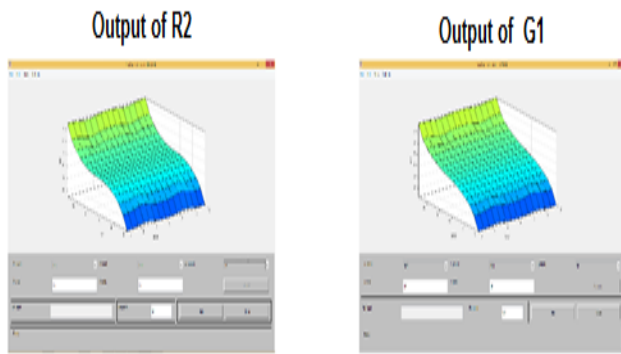


Fig. 7. Output of the Read Signal of Rod Two and Green Signal of Road One Both are Identical.

V. CONCLUSION

The system proposed here is very flexible. The fuzzy logic controller inputs can be taken from images taken from cameras fixed at each road. Due to the flexibility of fuzzy logic in dealing with stochastic systems, the traffic light control system showed good results in our simulations. The proposed FLSC is operated properly and efficiently in MATLAB environment. This controller gives a suitable green flexible timing that depends on the number of cars in each road, which can ensure vehicles are not allowed to wait too long on the road. While in the case of fixed time controller green time cannot be changed automatically, our system will give the green time according to the traffic congestion. So, arriving cars must activate the traffic signal time. The performance of the FLSC is affected by traffic congestion. The fuzzy logic control system has proven better in terms of improving the flow of traffic lights in the cities of Madinah and Makkah cities. In the future, this system will be proven on datasets from other countries, and to provide better results it will include a tracking stage to determine the traffic situation on the road before arriving at the traffic light is reached.

REFERENCES

- [1] R. Yuliani Kartikasari, G. Prakarsa, and D. Pradeka, "Optimization of Traffic Light Control Using Fuzzy Logic Sugeno Method", *International Journal of Global Operations Research*, Vol. 1, No. 2, pp. 51-61, 2020.
- [2] B. Ghazal, K. Elkhatab, K. Chahine, M. Kherfan, "Smart Traffic Light Control System", ISBN: 978-1-4673-6941-1 IEEE, 2016.
- [3] Roxanne Hawi, George Okeye, Michael Kimwele, "Smart traffic light control using fuzzy logic and wireless sensor network", *Computing Conference, London, UK*, INSPEC Accession Number: 17486620, DOI: 10.1109/SAI.2017.8252137. IEEE, 2017.
- [4] Kasun N. Hewage, Janaka Y. Ruwanpura, "Optimization of Traffic Light Timing Using Simulation", *Proceedings of the Winter Simulation Conference*, DOI: 10.1109/WSC.2004.1371482, 2004.
- [5] Nathanie Fairfield, Chris Urmson, "Traffic Light Mapping and Detection", *IEEE International Conference on Robotics and Automation*, 2011.
- [6] Sweta Pandey, Pratistha Mathur, Tejashri Patil, "Real time traffic signal control using fuzzy logic controller", *1st International Conference on Intelligent Systems and Information Management (ICISIM) INSPEC Accession Number: 17411069*, DOI: 10.1109/ICISIM.2017.8122190, IEEE, 2017.
- [7] J. Niityamki; V. Kononen, "traffic light controller based on fuzzy logic", *Smc conference proceeding iee. IAN.6778203*, IEEE. Nashville, TN, USA. 8-11 Oct. 2000.
- [8] V. Iyer; R. Jadhav, U. Mavchi, J. Abraham, "Intelligent traffic signal synchronization using fuzzy logic and Q-learning", *International Conference on Computing, Analytics and Security Trends (CAST)*, Pune, India, 2016.
- [9] Sandeep Mehan, "Introduction of Traffic Light Controller with Fuzzy Control System", *Dept. of ECE, RIEIT, Railmajra, Punjab, India*, ISSN: 2230-7109(Online) | ISSN: 2230-9543(Print), IEEE 2011.
- [10] Yi Hu, CQU, Peter Thomas, Member, IEEE, and Russel J. Stonier, Member, IEEE. "Traffic Signal Control using Fuzzy Logic and Evolutionary Algorithms". IEEE, 2007.
- [11] M. B. Jensen, M. P. Philipsen, A. Møgelmoose, T. B. Moeslund, and M. M. Trivedi, "Vision for Looking at Traffic Lights", pp 1800-1815, *IEEE Transactions on Intelligent Transportation Systems*, 2016.
- [12] Roul De Charette, Fawzi Nashashibi, "Traffic light recognition using image processing compared to learning processes", pp. 333-338, *IROS'09, IEEE*, 2009.
- [13] S. K. Kwon, E. Hyun, J.-H. Lee, J. Lee, and S. H. Son, "Detection scheme for a partially occluded pedestrian based on occluded depth in lidar radar sensor fusion", *Optical Engineering*, vol. 56, p. 1, 2017.

Private LTE Network Service Management Model, based on Agile Methodologies for Big Mining Companies

José Valdivia-Bedregal¹, Norka Bedregal-Alpaca², Elisa Castañeda-Huaman³
Universidad Nacional de San Agustín
Arequipa, Perú

Abstract—Information technology (IT) services must generate added value for any business, either by enhancing its processes, automating its activities, or managing its resources. Using Long Term Evolution (LTE) Networks, IoT solutions and devices are able to be deployed in order to prevent safety incidents and or to online monitor performance and maintenance indicators produced by field equipment. Under the same context, implementing a private LTE network Service Management Model becomes a basic need for any company. Proposed Service Management model must be flexible enough to changes in order to accomplish high productivity demands like the ones that a Big Mining Company requires. Proposed model is based on Information Technology Infrastructure Library (ITIL) as the best known, disseminated and proven framework; additionally, it uses well known agile methodologies such as Scrum and DevOps. Along with the deployment for each of the proposed stages, a visual scheme is generated which, when it comes to a conclusion at the final stage, allows to visualize model interactions in its entirety. In addition to describing the expected results, model validation has been accomplished by an expert panel judgment under the developed topic. As a conclusion, proposed model involves a holistic approach, that is, a comprehensive approach that addresses various aspects of service management supported by a private LTE platform.

Keywords—IT service management; agile methodology; ITIL; expert judgment; LTE network

I. INTRODUCTION

We are living through a permanent digital revolution and the so-called digital transformation has arrived to absolutely change everything; however, many organizations have not aligned their business strategy along with IT strategy in such a way it generates value for the business. This paradigm change, this migration to a new digital world, needs to consider new ways to work; traditional verticals and un-collaborative forms have not been always successful.

The IT area is responsible for aligning its strategic objectives with those defined by the business, being able to support business performance and regulatory indicators, technical and commercial demands within a flexible and change-adaptive framework.

Thus, [1] analyzes and combines the main characteristics of IT and network management models and frameworks, with a customer focus and process based. [2] develops an IT service management model applying agile frameworks and best

practice guidelines within IT management processes. At [3], the authors design and describe an information technology (IT) service management model, based on Cobit V5 framework and ITIL 2011 best practices. The author in [4] presents a comprehensive methodology for IT risk management based on ISO 31000 and ISO/IEC 27005 standards. At [5], a unified Medical and supplies management system is deployed by using agile SCRUM methodology.

In particular, big mining companies need to adapt their business strategies and models to be competitive enough. What used to be a physical focused activity and then focused at large machinery, is now moving towards knowledge and technology deployment. Under this context, implementing a private LTE network at a big mining company in addition to connecting mobile equipment will serve to remotely carry out many fundamental activities and also enable newer technologies such as augmented reality, digital twins, internet of things, among others. To achieve this objective, it is necessary to optimize Private LTE Service Management model which means a cultural change, involving change management and change adaptation; innovation and collaborative, agile and valuable efforts.

Under the described context, this article proposes a private LTE Service management model based on Agile Methodologies to improve productivity achievements according to a Big Mining company needs. In order to accomplish this, Section II displays theoretical sustentation, and Section III summarizes methodological study design. Section IV is the main part since it describes the aspects considered under ITIL framework and displays the general scheme for the proposal. The scheme allows the reader to see interrelationships between different implementation stages. Under Section V, expected results can be stated. Finally, to get a complete research validation, an expert panel with IT management experience and agile methodologies application was interviewed. Results are shown at Section VI.

II. CONCEPTUAL FRAMEWORK

A. Digital Transformation in Mining

Digital Transformation is a complex phenomenon in which five dimensions are identified: (a) digital leadership, (b) vision and digitalization strategy, (c) working methods, people and digital culture, (d) process digitization and decision-making, and (e) technology, data management and digital tools.

The question is how to effectively respond to the growing society digitization, not only in terms of how to avoid becoming obsolete along competition, but also how to adapt and lead digital disruption [6].

With the fourth industrial revolution, mining operations will be marked by the use of Cyber Physical Systems, Artificial Intelligence, Internet of Things (IoT) and Big Data, for process optimization. ICT solutions applied to the mining industry define a new idea an intelligent mine, for the Mining Skills Council [7] three facets must be considered: mechanization, telecommunication-automation and optimization.

B. Technology Solution: Long Term Evolution (LTE)

Long Term Evolution, known by its acronym LTE, is a high-capacity radio technology standardized by 3GPP. LTE is a stable technology with three main features: it allows high bit rates with low latency, is cheap and easy to deploy by operators, avoids fragmentation by duplication type. LTE-M is a type of network designed to deploy a way to communicate machines and robots beyond sensors, it is characterized because it can support a wide spectrum of M2M devices with low cost and long battery life.

C. Information Technology Services Management (ITSM)

ITSM is a strategic approach to design, delivery, management and improvement around ICTs within an organization. For [8], good IT service management should aim to: provide adequate quality management, increase efficiency, align business processes and ICT infrastructure, reduce risks associated with IT services and generate business value.

IT Services Management frameworks include Six Sigma, Microsoft Operations Framework, COBIT, ITIL (Information Technology Infrastructure Library).

III. METHODOLOGY

The article can be considered as applied research. It focuses on proposing a model to address a specific problem and achieve a specific objective.

By the level of depth, it is an exploratory – descriptive research, because it analyzes and investigates specific reality aspects that have not yet been analyzed in depth and establishes a description of a specific context.

Study units are the IT strategies as business strategies support related to the three main indicators in big mining: production, cost and safety.

Data collection techniques used are: observation, interview and documentary research.

IV. PROPOSED MODEL

The following is a description of an stepped process to be followed by an “La Minera” IT department to adopt an agile Private LTE Network Service Management framework. The following stages have been considered: strategic, design, transition, operation and transversal processes.

Fig. 1 outlines the proposal for each of the stages considered.

A. Strategic Stage Development

At this stage, strategies are defined considering that they must be flexible enough to adapt quickly to core business strategy and maintain or achieve competitive advantages.

In relation to "Financial Management", it is proposed that management and maintenance costs must be considered at department's annual budget. To this end, annual budget has been defined for 4 years with 2 initial contracts, first one for an on-site level 1 support and the second one for a remote level 2 support provided by the telecommunications main operator. The sub-processes considered are (a) Financial Management Support, (b) Financial Planning (Budget Definition), (c) Financial Analysis and Reporting (Accounting) and (d) Service Billing (Collections).

The proposed strategy for "Business Relations" considers establishing and maintaining a client focused business relationship based on understanding the client and their needs, for which the Business Cases must be early defined and then implemented over the LTE platform. Currently under process there is: Heavy fleet management improvement, heavy fleet video solution migration, and implementing a unified geotechnical system for slope and water monitoring.

"Demand Management" considers that the implemented private LTE platform has an initial 95% coverage capacity within the operation and a bandwidth that supports 40 MB/s, therefore some limitations must be analyzed later during project implementation such as the platform being able to connect only 1000 concurrent clients due to licensing or lack of band aggregation for capacity improvement.

In relation to "Service Portfolio Management", in "La Minera", initial investment made for the LTE platform can only be evaluated when reviewing the services that depend on it or that have been improved by the platform. At the moment, it is expected to increase the reliability and management capacity of the company's heavy fleet. As more services are deployed over the private LTE platform, it is relevant to perform periodical capacity reviews to analyze if services deployed over the platform are really needed or should be considered for decommission.

B. Design Stage Development

At this stage, both the design of new services and changes and improvements to existing ones must be considered; all this in order to meet the present and future requirements of the company.

In relation to "Supplier Management" three levels of support have been considered for service management. Level 1 support is provided by a company that will be in charge of all on-site support and will be responsible for monitoring and giving the first quick response to all the platform equipment and to the high criticality CPEs. Level 2 Support, which is provided by the company that rents the Band 700; this provider will be in charge of solving the problems that, due to lack of technical specialization Level 1 cannot solve. Level 3 Support is provided by the hardware and solution providers, i.e., Nokia, Cisco and Italtel. This set of support levels theoretically ensures 99.5% availability for the platform.

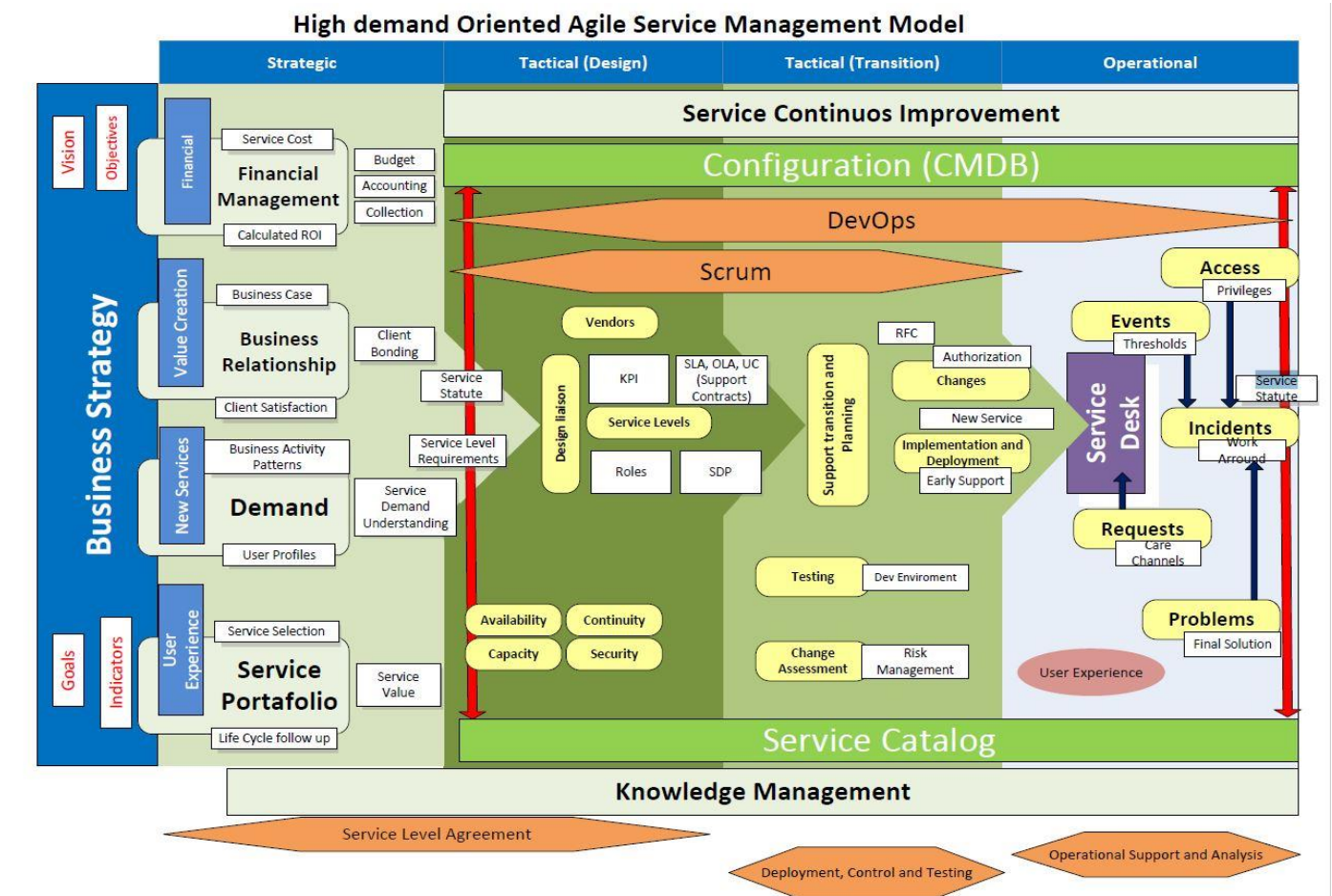


Fig. 1. General Scheme of the Proposed Model.

The "Coordination of Service Design" becomes important for the transition of new services to be hung on the platform. Service Design Packages (SDP) will be specified to assist in the delivery of the design to the transition and operation. They will primarily be used to set the parameters for future services that hang from the LTE platform.

For "Definition of Service Levels" information security is considered important when deploying an Internet of Things (IoT) platform such as the one being used on the LTE network, therefore, parameters for the platform must be defined in advance: LTE Platform Availability, LTE Platform Capacity, LTE Platform Continuity, and LTE Platform Security.

C. Transition Stage Development

The objective of this stage is to ensure the use of procedures that ensure the efficiency and quality of each planned change, the realization and integration of new services should take place without disturbing the operation and performance of existing services.

With regard to "Transition Planning and Support", procedures for successful deployment of new services or changes on the LTE platform must be defined. Any changes must be registered in the IT Service management platform. Two important aspects have been considered: Change Management and the Deployment and Implementation

Procedure. Change Management takes priority care of the authorizations and permissions necessary to work on the platform and reduce a possible impact, considers three levels of approval: Technical, Change CAB and Business.

The "Deployment and Implementation" process defines an early support scheme so that the area of operation is in a better position to operate the service.

For the "Validation and Testing" process, the infrastructure was installed and a haul truck was determined to start testing: the HT110 truck. This was used on the same Dispatch System that is used for the operation. Initially two PTX computer interaction screens were installed, each corresponding to one of the Connection systems: LTE and Mesh Wi-Fi. As time went by and the tests were fine-tuned, only the PTX corresponding to the LTE network was continued to use. As a result, the LTE network was faster, more reliable and with greater coverage than the one using the old Mesh Wi-Fi network.

D. Operation Stage Development

Ensuring that IT services are offered effectively and efficiently must meet user requirements, resolve service failures, troubleshoot issues, and perform routine operations.

For "Access Management" two types have been defined: the first at the administrator level, for the people who will manage the service and the second at the user level, which will

be the customers of the Private LTE platform. At the administrator level, three user levels must be created. At level 1 they will have full access to all systems; all the experts in the organization are included in this group. In Level 2 and Level 3 Core Support, Italtel and Cisco teams are considered. RAN Support Level 2 and Level 3 consider the Nokia team. The user level must be operated at two levels, depending on the type of network they will connect to and the service being used.

"Event Management" focuses on the correct managing of the Network Monitoring System (NMS). The SolarWinds license is available for the corporate and industrial networks. Considering the good experiences in the use of this license; a hybrid solution of SolarWinds with Cisco Prime is being purchased for the LTE network. During the implementation of the NMS, operation thresholds must be defined. For example, client computers (CPEs) should not exceed 85% CPU or memory usage. It is necessary to define the usage threshold that each of the APNs will have depending on the available capacity, in this case, the events must be triggered at 80% capacity usage.

"Incident Management" must provide a definitive or temporary solution to any incident that may occurs. With the confidence of having a 24/7 monitoring center, every incident will be recorded with a unique identifier in the ITSM of "La Minera". It is expected that most incidents will be logged by the NOC (Network Operation Centre) monitoring center, ensuring a proactive rather than reactive style of working. Any user can log an incident to the Service Desk considering two types: platform incidents or client incidents.

The "Request Management" as a regular process begins with the review of the application by the Digital Transformation area, who will evaluate and classify the activity in two main groups: common tasks for the business and initiatives to be consider by the Road Map.

"Problem Management" must prevent problems and incidents with the LTE network, it is possible that some customers lose connection in an area constantly, then it is part of problem management to identify the pattern of problem and seek a definitive solution, so that the same incident does not continue to occur. It is important to investigate incidents of high impact and high priority, the results of the investigation will be recorded in a Post Incident Review (PIR) document in which the root cause and corrective actions in the medium term are defined.

As a result of the operation of the service there will be multiple opinions of the business, which will depend on their past and present experiences in relation to the LTE platform and the services that work on it. For the "User Experience Management", the Service Desk will have the responsibility to implement surveys on the attention received and to provide a platform to record initiatives for potential applications or new services that could run on the platform. All this information must be registered with the ITSM: Service Now.

E. Transversal to All Stage Processes Development

Cross-cutting processes are those that interact and are relevant to all the stages described above, breaking the vertical activity flow scheme and managing to encompass the entire

proposed structure. Three processes have been considered: Knowledge Management, Configuration Management, and Continuous Improvement of Services. The DevOps and Scrum methodologies will be used in the implementation of new services on the platform, focusing on not creating any affectation to the already implemented services. Both methodologies consider design and definition moments that must be addressed in order to generate good practices in the application of these methodologies.

V. EXPECTED RESULTS

Generating a strategy to manage the LTE platform in order to meet the needs of "La Minera" allowed to determine processes to meet the needs of the business without losing control and correct management of the new network and identify a set of benefits derived from the use of the LTE platform and the proposed model.

The cost of implementing the projects already defined in the Road Map of the area and of the company to implement the digital transformation will be reduced. For example, the Unified Slope Monitoring project for the Geotechnical area reduced its cost by approximately 16%, which is very significant, since the maintenance of the slopes represents one of the eight Fatal Risks in the business.

It will be possible to increase the volume of production, for which a project (supported by the LTE network) has already been registered, which considers the implementation of VisioFrothTM, that is, an advanced image analysis system that monitors in real time a wide variety of characteristics of the flotation cells, characteristics that can be used to adjust the level of the cells or the feed air flow, and thus guarantee maximum mineral recovery in the flotation stage.

Security risks will be reduced, as the first services already implemented on the LTE network are the Unified Slope Monitoring System and the Electric Storm Monitoring and Alert System. Both systems have a stable, easily accessible network, with the possibility of mobile positions in new locations without involving the placement of links or network cabling that complicate the work.

The risks of production losses will also be reduced, the control of the mills supported in the LTE network will prevent unplanned falls of equipment in operation, reducing maintenance times considerably and favoring the objective of ensuring the availability of SAG mills to 99.9%.

It will be completed in the appropriate time with the report to entities such as the National Water Authority or the Environmental Evaluation and Control Agency, reports that involve information related to sampling. Currently the process is manual; however, taking advantage of LTE network connectivity, automatic monitoring equipment can be installed for sampling. On the other hand, with large volumes of data from automated monitoring systems, data mining processes could be implemented to establish patterns and predict future events [8].

By implementing the best practices proposed by the ITIL framework, it is expected to improve the management IT area's development and production environments; standardizing

processes with the standard documentation indicated by ITIL will greatly improve transitions between life cycle stages and reduce early project support and incident response times.

With the application of SCRUM, implementation flexibility will be increased by having a Dynamic Backlog; each time someone wants to bring about a change in the scope of the project, a new item is added to the backlog, without having to meet with the Steering Committee. SCRUM allows the project to turn according to the real and changing needs of the business, it is expected to reduce the administrative time in the various changes of the scope of the project to a maximum of nine days which is what each Sprint will last and thereby reduce the delays by approximately 40%.

By using DevOPS best practices, it is ensured that change management is strictly controlled, thus avoiding potential business impacts. Documentation and implementation processes become pillars for any implemented changes, helping to achieve the goal of eliminating known risks and reducing the likelihood of unknown risk during implementation.

VI. PROPOSAL VALIDATION

To validate the usefulness of the proposed model, a panel of experts with experience in IT management and the application of agile methodologies were used. A professional biography of the expert was developed based on his or her background: years of experience and training, research or training actions, knowledge of the subject matter of study. For [9], the number of experts depends on the ease of accessing them or the possibility of meeting sufficient experts on the subject matter of the research, so nine experts were chosen.

The Expert Competence Coefficient (k) was calculated as the average of two quantitative indicators the knowledge coefficient (kc) and the argumentation coefficient (ka). The kc coefficient is a measure of the level of knowledge about the investigated topic that comes from the self-assessment of each candidate. The ka coefficient was obtained from assessing its professional merits: related research, work experience, specialized reading, and knowledge of the problem situation, specialization or postgraduate studies, alignment with best practices in the sector, alignment with the business model and vision of the future. The results obtained are shown in Table I.

Of the nine experts selected, seven of them are in a high degree of competence, so it could be said that the results show high competence of the experts.

The definition of the criteria that served to build the questionnaire that would apply to the experts was based on the work of [10] and [11] who present a list of benefits related to the implementation of ITIL. Each item of the questionnaire was rated on a 5-level Likert-type scale. The Likert scale is a measurement tool that allows to measure attitudes and to know the degree of conformity of the respondent with the statements proposed to him/her [12].

Table II shows the percentage of opinions that fall within each level for each of the questions asked in the questionnaire; the last column contains the average score obtained on the item.

When calculating the average of the scores obtained for the 13 items, a value of 4.25 is obtained (red line in Fig. 2); a value that is well above the scale average of 3 points.

It can be seen in Fig. 2 that 7 of the 13 questions exceeded the average (4.25), three of them scored very close and 3 were far from the average. However, the least valued item (item 13) was the only one that was slightly below the scale average. It can be concluded then that the overall assessment of the experts is positive and in agreement with the proposal.

From the evaluations made by the experts, it is clear that they agree that the proposed model will have an impact (high or very high) on the improvement of the quality of the service delivered to the customer considering their specific needs, which will increase customer satisfaction; it will also have a strong impact on the management of IT operations as a service.

In other words, an adequate IT Service Management (ITSM) would be carried out to ensure that IT objectives are aligned with business objectives. In addition, the standardization of processes and services would be achieved through the definition of functions and documented processes for the delivery of each type of IT service.

As a measure of the reliability of the internal consistency of the items, Cronbach's α coefficient was calculated. An $\alpha = 0.78$ was obtained, so it can be said that the evaluation made by the experts has a "Good" reliability, close to "High".

Additionally, the experts were asked to identify success factors in the implementation of the model.

TABLE I. DETERMINATION OF THE K COEFFICIENT AND THE LEVEL OF COMPETENCE OF EACH EXPERT

Expert	kc	ka	k	Competence level
E-01	0.8	0.89	0.845	High
E-02	0.9	0.9	0.9	High
E-03	1	0.91	0.955	High
E-04	1	0.96	0.98	High
E-05	0.8	0.9	0.85	High
E-06	0.8	0.76	0.78	Medium
E-07	0.9	0.84	0.87	High
E-08	0.7	0.82	0.76	Medium
E-09	0.9	0.95	0.925	High

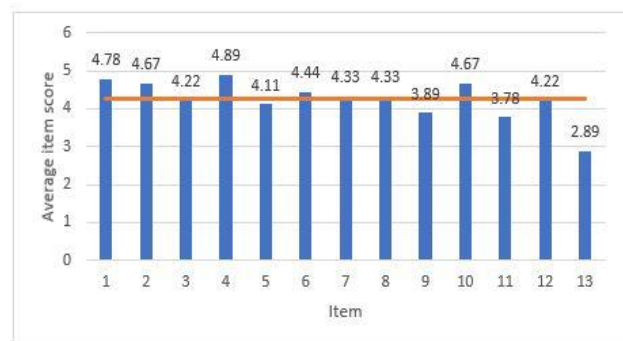


Fig. 2. Average Scores Obtained for Each Item.

TABLE II. PERCENTAGE OF OPINIONS IN FAVOR OF EACH LEVEL

Item	(1) Very low impact	(2) Low impact	(3) Moderate impact	(4) High impact	(5) Very high impact	Average
Improve quality and level of service	0	0	0	22.22	77.78	4.78
Increase customer satisfaction	0	0	0	33.33	66.67	4.67
Standardized and more effective processes	0	0	0	77.78	22.22	4.22
Continuous Service Improvement	0	0	0	11.11	88.89	4.89
Improve IT interaction with the rest of the business.	0	0	0	88.89	11.11	4.11
Adopt a common IT process methodology	0	0	11.11	33.33	55.56	4.44
Alignment of services, processes and goals with the requirements of the organization	0	0	11.11	44.44	44.45	4.33
Reduce downtime in IT services	0	0	11.11	44.45	44.44	4.33
Cost reduction	0	0	33.33	44.44	22.23	3.89
Improve company productivity	0	0	0	33.33	66.67	4.67
Competitive advantage over other suppliers	0	0	44.44	33.33	22.23	3.78
Increasing IT predictability and efficiency	0	0	11.11	55.56	33.33	4.22
Comply with regulations	0	33.34	44.44	22.22	0	2.89

The experts consider that the factors: Senior Management Support, Staff Training and Development, Managing the cultural change from a technological approach to a service-centered approach and Strong focus on continuous improvement will have a strong influence (quite influential and highly influential) on the success of the model implementation (Fig. 3, 4, 5 and 6).

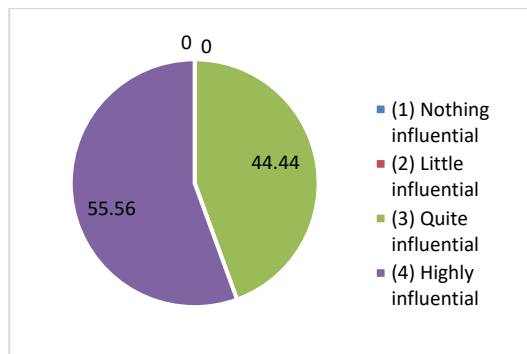


Fig. 3. Senior Management Support.

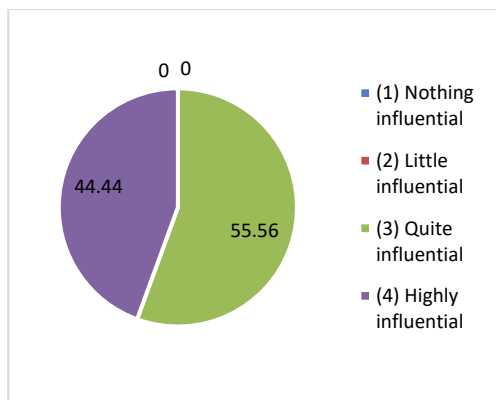


Fig. 4. Staff Training and Development.

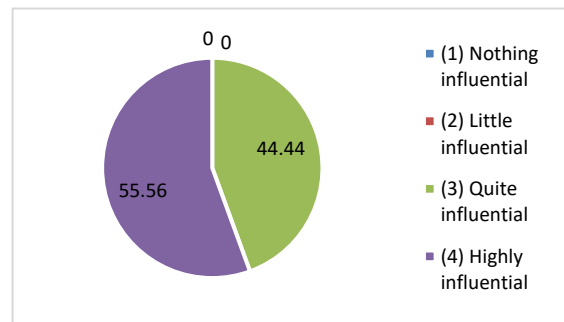


Fig. 5. Managing Cultural Change from a Technology-Driven Approach to a Service-Focused Approach.

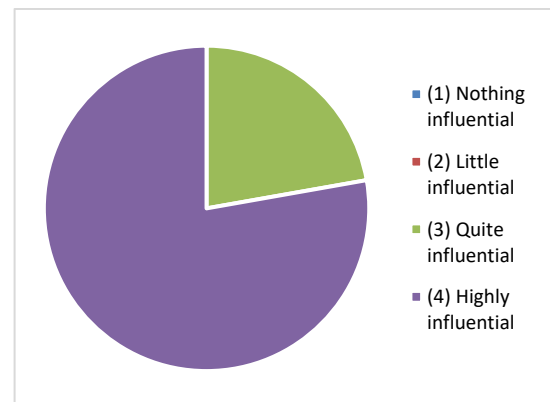


Fig. 6. Strong Focus on Continuous Improvement.

VII. CONCLUSIONS

Considering the rapidly changing company needs, especially at the mining sector, IT functions must increasingly be integrated into day-to-day operations as in other sectors like banking. It is appropriate to adopt IT Services Management (ITSM) best practices, including the ITIL framework.

A private LTE network Service Management Model is necessary, since the digital transformation of "La Minera" mining not only considers information systems interoperability between operation and production, but also includes important process digitalization of the mining operation environment such as: optimizing and monitoring mobile mine fleet conditions, driver behavior monitoring, anti-collision systems, IoT technologies deployment and integration, ore tracking, fleet automation, among others.

To perform what it is called digital mining, it is necessary to keep all information systems connected; as much information is available online, better informed decisions can be made; under this context, an adequate service management supported by the LTE platform is not only an IT strategy, but it becomes a business strategy.

It was possible to identify and describe various models, standards and frameworks, related to IT service management, including ITIL, Scrum, DevOps. In addition, their applicability and compatibility were identified under the context of managing services supported by LTE platform through the proposed model.

Considering that ITIL is a predictive framework of good practices at an IT service management department, which is usually applied under well-known scopes, that Scrum is an agile project management methodology oriented to less known scopes and that DevOps is a process that emphasizes communication and collaboration seeking to reduce change risk, a holistic approach model has been proposed, i.e. a comprehensive approach that addresses various aspects of service management for LTE platform supported services.

ACKNOWLEDGMENT

The authors express their appreciation to the National University of San Augustin of Arequipa for being the study center that allowed the realization of this research work.

REFERENCES

[1] R. Padilla and M. Ron, "Propuesta de modelo de gestión de infraestructura de red, basado en las mejores prácticas de gestión de TI y

los modelos estándar de gestión de red- caso de estudio EP PETROECUADOR", Tesis de maestría en gestión de las comunicaciones, Escuela politécnica nacional, Quito, Ecuador, 2015.

[2] C. Navarrete and M. Ramos, "Modelo de gestión de servicios de TI para la Organización Panamericana de la Salud en el Ecuador", Tesis de maestría en gestión de las comunicaciones, Escuela politécnica nacional, Quito, Ecuador, 2016.

[3] D. López, "Modelo de gestión de los servicios de tecnología de información basado en COBIT, ITIL e ISO/IEC 27000", Revista Tecnológica - ESPOL, 30(1), 2017.

[4] F. Arévalo, I. Cedillo and S. Moscoso, "Metodología Ágil para la Gestión de Riesgos Informáticos", Revista Killkana Técnica. Vol. 1, No. 2, pp. 31-42, 2017. p-ISSN 2528-8024 / e-ISSN 2588-0888, Universidad Católica de Cuenca, 2017.

[5] H. Robalino, E. Rodríguez and A. Saldaña, "Diseño de un sistema de gestión de servicios aplicando las buenas prácticas ITIL 2011 y SCRUM en el área de Soporte de Sistemas para la empresa APC Corporación", Universidad Tecnológica del Perú, 2019.

[6] E. Schreckling and C. Steiger, "Digitalize or Drown, In book: Shaping the Digital Enterprise", August 2017. DOI: 10.1007/978-3-319-40967-2_1.

[7] Consejo de Competencias Mineras, "Impacto de las nuevas tecnologías en las competencias requeridas por la industria minera", Alder Comunicaciones, Santiago, Chile, 2018.

[8] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, P. Yanque-Churo, "Classification models for determining types of academic risk and predicting dropout in university students". International Journal of Advanced Computer Science and Applications (IJACSA), Volume 11 Issue 1, 2020. DOI 10.14569/IJACSA.2020.0110133.

[9] J. Cabero Almenara and M. Llorente Cejudo, "La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información (TIC)". En Eduweb, Revista de Tecnología de Información y Comunicación en Educación, 7 (2) pp.11-22, 2013.

[10] N. Lucio and D. Gonzalez-Bañales, "Prácticas de ITSM en México y Latinoamérica 2014", México: PITSM Latam, 2014.

[11] Kumbakara, Narayanan, "Managed IT services: the role of IT standards", Information Management & Computer Security, Vol. 16 Iss: 4, pp.336 – 359, 2008.

[12] N. Bedregal-Alpaca, V. Cornejo-Aparicio, D. Tupacyupanqui-Jaén and S. Flores-Silva, "Evaluación de la percepción estudiantil en relación al uso de la plataforma Moodle desde la perspectiva del TAM", Ingeniare. Revista chilena de ingeniería versión On-line ISSN 0718-3305, Ingeniare. Rev. chil. ing. vol.27 no.4 Arica dic. 2019. DOI 10.4067/S0718-33052019000400707.

Speeding up Natural Language Text Search using Compression

Majed AbuSafiya

Software Engineering Department
Al-Ahliyya Amman University
Amman, Jordan

Abstract—Text search is a well-known problem in computer science where the valid shifts of a pattern P in a text string T are found. This paper shows how to speed up text search by searching for P in a compressed version of T . A fast compression algorithm was designed for this aim. This algorithm is based on the assumption that T is restricted to the letters of a single natural language. Relying on this assumption, a letter, in T or P , is encoded into a single byte instead of the two-byte unicode which shortens the string on which a text search algorithm works. The main disadvantage of this approach is the restriction of the alphabet of T to be from a single natural language. However, wide range of text documents complies to this assumption. Another issue is the overhead that is required to compress P and T , but it was found that the proposed compression algorithm is so fast such that its run-time can be paid for and still save text search time. Different approaches to store compressed T are also explored. The conducted experimental study showed that this approach does actually reduce the text search time.

Keywords—Text compression; text search; unicode

I. INTRODUCTION

A lot of research was directed towards searching in compressed text. A survey of the approaches to search in compressed text without decompression can be found in [1]. In [2], text search was applied on a directory-based compressed text. In [3], the characters are encoded as a variable-length sequences of base symbols of fixed number of bits. In [4], the input text is already compressed with Lempel-Ziv. In [5], a compression and decompression techniques for natural language text are proposed. The compression scheme that is used is based on semi-static word-based model and Huffman encoding where the coded alphabet is byte oriented rather than bit-oriented. In [6], an approximate search on the compressed search using local decompression is proposed. In [7] the input text is assumed to be Ziv-Lempel Compressed Text. In [8], the text search in compressed text is done using periodicity analysis, with sublinear run time with the size of compressed text. In [9], a directory based compression is used on natural language text.

The main observation that can be noticed in the previous research regarding this problem is that: the primary motivation was to do text search in an input string that is already compressed using known compression algorithms without decompressing it first. This means that compression was not originally done to speed text search. This is the main point that contrasts this work from others work. In this paper,

compression is done to speed up the text search first and to save space as a second gain. The compression that was considered in literature is based on known compression algorithms which are known to be complex and time-consuming. On the other hand, the proposed compression is very fast and simple. A similar approach to our approach was found in [10]. However this work differs from our work in many ways: (1) T is assumed to be in ASCII while our work is based on unicode encoding, (2) our compression approach is much faster and simpler, (3) the shifts that are found in the compressed T can easily be translated to shifts in the original unencoded version of T .

The proposed work in this paper is based on the fact that T is encoded in unicode [11]. Unicode is an international standard for encoding alphabets of natural languages, two bytes for a letter. Alphabets of different natural languages are encoded in ranges. One observation about unicode is that the alphabet of the same natural language share the same upper byte value. For example, Arabic alphabet unicodes range between 0x0600 up to 0x6FF with the same upper byte code 0x60. This fact will be utilized to compress T and P to reduce their length to half by excluding the upper byte. For example, the Arabic word (هو) is composed of two letters (Fig. 1). The unicode of letter (هـ) is 0x0647, the upper byte is 0x06 and the lower byte is 0x47. The unicode of the second letter (و) is 0x0648 with upper byte is 0x06 and lower byte is 0x48. The proposed compression is based on using only the lower byte as a code for the letter. This will compress T into half size. Note that this compression works only under the assumption that T contains only text letters of the same alphabet. Moreover, the code of the first letter (هـ) is placed in the upper byte in the compressed unicode. This is because Arabic script is written from right to left and hence it assures that the encoded letters will be stored in the same order that they have within the original text.

هـ و		Word
0x0647	0x0648	unicode
0x4748		Compressed code

Fig. 1. Compression of Two Letters into One Letter.

II. COMPRESSION ALGORITHM

A. COMPRESS Algorithm

COMPRESS algorithm (Fig. 2) returns a compressed string (*Scompressed*) for an input unicode string S . S could be T or P .

$S_{compressed}$ will be half the length of S . S is a string, where each letter is encoded with two-byte unicode. $S_{compressed}$ is generated by reducing the two-byte unicode code of every letter in S into a single byte in $S_{compressed}$ (Fig. 1). To show how this algorithm works for the example in Fig. 1, let S be the string (هو) with two letters S_0 =(ه) and S_1 =(و). The algorithm will do a left-shift on S_0 by eight bits to generate S'_0 (0x4700). Next, a bit-wise *and* operation is applied between S_1 and 0x00FF to generate S'_1 (0x0048). Finally, a bit-wise *or* operation is applied between S'_0 and S'_1 which will result in one letter unicode (i.e. 0x4748) that will be appended to $S_{compressed}$. The loop will iterate for the letters of S in pairs until $S_{compressed}$ is complete. In case of S is of odd length, a *space* will be appended to make its length even.

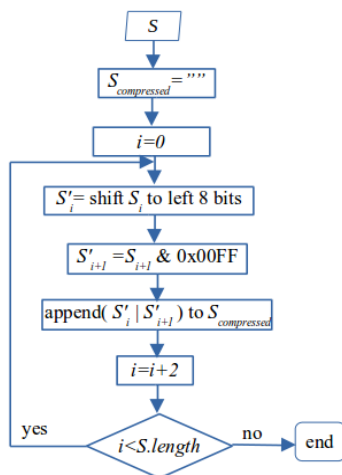


Fig. 2. COMPRESS Algorithm.

B. Saving $T_{compressed}$ to a file

Saving $T_{compressed}$ into a file has two advantages: (1) saving disk space, if it replaces the original T 's text file since the compressed version is half the size of the T 's text file, (2) allowing immediate application of the text search on the compressed version of T saves the time to compress T every time a text search is required. It is important to point here that the experimental study showed reduction is search time even if T is input in its native uncompressed format and compression is done as part of the text search. To save T in a compressed form, the text file of T is read and COMPRESS algorithm is called to build $T_{compressed}$. Remember that $T_{compressed}$ is an array of bytes, one byte encodes one letter in T . There are two approaches to write $T_{compressed}$: as a unicode text file or as a binary file.

One way to store $T_{compressed}$ is through storing it in a text file using unicode. In this case, every *pair* of bytes of $T_{compressed}$ is interpreted as a single unicode code character

and the corresponding character of this unicode is written to the file. So, the size of this file will be half the size of the original T 's file. In addition to compression, it will be encrypted. For example, T = “بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ”, which is composed of twenty two characters (letters and spaces), will be encoded into eleven unicode characters. Fig. 3 shows how $T_{compressed}$ looks like when its file is opened in a text editor.

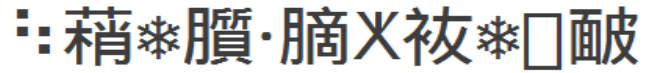


Fig. 3. $T_{compressed}$ as Unicode Text.

$T_{compressed}$ may also be stored as binary file (Fig. 4). The letters of $T_{compressed}$ are written into a binary file a byte by byte without building unicode letters from pairs of bytes. In this case, the letter is represented as ASCII code. For example, for T =“بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ”, each character is represented by one byte. This byte corresponds to the lower byte of the unicode of that character. For example the first byte is 'ب' which has the hex value 0x28 is the code of the first letter (ب). Note that the unicode for letter (ب) is 0x0628.



Fig. 4. $T_{compressed}$ Stored as Binary.

III. SPEEDING TEXT SEARCH WITH COMPRESSION

The compressed text search algorithm is shown in Fig. 5. It takes as input P and T , compresses them using COMPRESS algorithm and then calls any known string matching algorithm to search for $P_{compressed}$ within $T_{compressed}$. Although $T_{compressed}$ has half the size of T , the length of $T_{compressed}$ equals to the length of T . This is because $T_{compressed}$ is viewed as an array of bytes (on byte for each letter) and T is viewed as an array of unicodes. This is also true for P and $P_{compressed}$. The equality comparisons within the selected string matching algorithm will be byte-wise comparisons and not unicodes comparisons. So, the calculated valid shifts that are found by the this reused string matching algorithm will be the valid shifts of P within T .

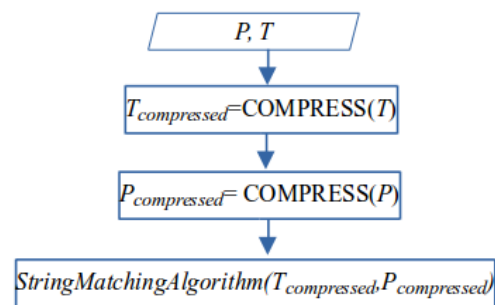


Fig. 5. Compressed Text Search Algorithm.

IV. EXPERIMENTAL STUDY

The compressed text search algorithm was implemented in Java, where T is chosen to be the text of the Holy Quran, which is composed only of Arabic letters, with size of 411,082 letters. The selected string matching algorithm was the known Knuth–Morris–Pratt (KMP) algorithm [12]. To show the reduction in text search time, the search time that is needed to search for P in T using the compressed text search algorithm is compared with the time that is needed to search for P in T without compression (Fig. 6). P was randomly chosen as a substring of a given length from T . This experiment was repeated for varying sizes of P . Note that the time to compress P and T was included in calculating the search time for the compressed text search algorithm. To raise the confidence in the results, this process was repeated 1000 times for each length of P and the average time was recorded for both algorithms. It is obvious that when KMP is applied on compressed input, it resulted in significant reduction in search time. The saving in time happened because an equality comparison between two unicode characters, in Java, is actually implemented through a couple of byte-wise equality comparisons. On the other hand, when searching in $T_{compressed}$, the equality comparison is done by a single byte-wise comparison.

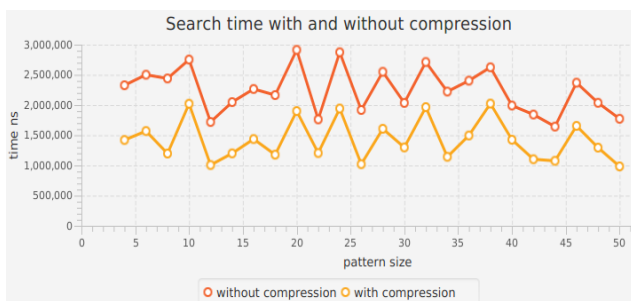


Fig. 6. Comparison between KMP Time with/without Compression of P and T .

V. CONCLUSIONS

In this paper, the natural language text was compressed to speed up text search. By excluding the upper byte of the unicode of letters, we could reduce the size of both P and T into half and hence have a faster text search. This approach assumes that letters of the text belong to the alphabet of the same natural language. One important result from this research is that exploiting the specifics and constraints of natural languages may open the door for improvements on string

algorithms in general. Although these improvements are not generic, they may be useful under certain contexts. One interesting issue to explore is how to do text search when $T_{compressed}$ and $P_{compressed}$ are viewed as arrays of unicodes rather than arrays of bytes. The challenge here is to explore how to compress P such that the odd valid shifts of P within T are also found.

ACKNOWLEDGMENT

All praise and gratitude be to *Allah*, all mighty, for guiding me and giving me the knowledge and strength to accomplish this work.

REFERENCES

- [1] D. Adjero, T. Bell, and A. Mukherjee, Pattern Matching in Compressed Texts and Images. Now Publishers Inc., Hanover, MA, 2013.
- [2] K. Fredriksson and S. Grabowski, "A general compression algorithm that supports fast searching," Information Processing Letters, vol. 100, 2006, pp.226-232.
- [3] J. Rautio, J. Tanninen and J. Tarhio, "String matching with stopper compression," Proceedings DCC 2002, Data Compression Conference, Snowbird, UT, USA, 2002, pp. 469-476.
- [4] M. Farach-Colton and M. Thorup, "String Matching in Lempel–Ziv Compressed Strings," Algorithmica, vol.20, 1998, pp. 388-404.
- [5] E. de Moura, G. Navarro, N. Ziviani, and R. Baeza-Yates, "Fast and flexible word searching on compressed text," ACM Trans. Inf. Syst. Vol. 18, 2000, pp. 113–139.
- [6] G. Navarro, T. Kida, M. Takeda, A. Shinohara and S. Arikawa, "Faster approximate string matching over compressed text," Proceedings DCC 2001. Data Compression Conference, Snowbird, UT, USA, 2001, pp. 459-468.
- [7] G. Navarro and J. Tarhio, "Boyer-Moore String Matching over Ziv-Lempel Compressed Text," Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching (COM '00). Springer-Verlag, Berlin, Heidelberg, pp. 166–180, 2000.
- [8] A. Amir and G. Benson, "Two-dimensional periodicity and its applications," In Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms (SODA '92). Society for Industrial and Applied Mathematics, USA, 1992, pp. 440–452.
- [9] K. Fredriksson and F. Nikitin, "Simple Compression Code Supporting Random Access and Fast String Matching," In: Demetrescu C. (eds) Experimental Algorithms. WEA 2007. Lecture Notes in Computer Science, vol 4525. Springer, Berlin, Heidelberg.
- [10] Udi Manber, "A text compression scheme that allows fast searching directly in the compressed file," ACM Trans. Inf. Syst. 15, pp. 124–136, 1997.[12] D. Knuth; J. Morris, V. Pratt, "Fast pattern matching in strings," SIAM Journal on Computing vol.6, 1977, pp. 323-350.
- [11] Unicode home, <http://home.unicode.org>.
- [12] D. Knuth; J. Morris, V. Pratt, "Fast pattern matching in strings," SIAM Journal on Computing vol.6, 1977, pp. 323-350.

Developing an IoT Platform for the Elderly Health Care

Medhat Awadalla¹, Firdous Kausar², Razzaqul Ahshan³

Department of Electrical and Computer Engineering
College of Engineering, Sultan Qaboos University
Muscat, Oman

Abstract—The health care of elderly people addresses the necessity for services that utilize recent technologies and devices. Now-a-days, both loneliness and psychological depressions are typical problems which elderly people face because of living alone/abandoned or reduced communication with their children and relatives. This paper presents the development of an integrated platform using the Internet of Things to manage and provide extensive services for elderly people to address the aforementioned issues. The proposed platform relies on wearable sensor devices to collect real-time data and store it in a cloud server via a developed smartphone application. The cloud server is accessed to retrieve data stored using the OAuth protocol. A web-based database-driven application is developed that facilitates the management of helpful information about elderly people to an authorized person. The doctors can perform real-time monitoring of the health condition of their elder patients remotely, and the system generates an alarm and sends notifications to caregivers and doctors in case of emergency. The conducted experiments and the achieved results showed that the developed platform has remarkable features of accessibility, security, efficiency, and cost.

Keywords—Internet of things; healthcare; sensor network; real-time monitoring; wearable sensors; wireless sensor networks

I. INTRODUCTION

Nowadays, the number of elderly is growing up, and it is expected to reach from 962 million globally in 2017 to 2.1 billion in 2050 [1]. Many elderly people lose their strength and ability to serve themselves if they are living alone. However, many of the elderly live alone in some societies since their children leave their homes when they reach the age of 18th. In Oman, the whole family lived together in the same village in the past because most of them worked in the same area, so there was no problem taking care of such issues. Nowadays, due to the lack of jobs in villages, people leave their villages and go to cities for jobs, which most probably be very away from their homes. Consequently, this will cause problems in monitoring their elderly health.

It is a global concern, and many of the elderly have numerous health problems such as hypertension and hypothyroidism, diabetes, heartbeat problems, and much more. Such illnesses need to be monitored to ensure no risk of death due to any sudden changes. In some countries, like Oman, it is impossible to leave the elderly to stay in some organizations like "Elderly Home" to monitor their health; since it is considered ingratitude, the same situation for all other Arabian countries. The hospitals cannot monitor all the elderly people

constantly inside their buildings due to the limitations of medical staff and beds. Furthermore, the escorts cannot regularly stay in the hospital since they have other work to do, especially if it is not an emergency case that requires the elderly people to stay at the hospital. Moreover, when people become old, their brains do not have the same ability as adults in remembering things, which means that they may forget how to get back to the house if they go outside. The relatives will face problems finding them, and the risk will increase if the elderly suffer from any diseases.

Therefore, a system is proposed which enables continuous monitoring for elderly people's health in real-time to prevent chronic diseases, thus preventing hospitalization that burden the healthcare systems and costs. The proposed system measures some parameters and monitors the elderly people at home or in a particular area outside the home. This system works automatically using wireless communication and also monitors heartbeat, steps, and calories. It gives an alert to the hospital and caregivers when a problem occurs. Also, the health data information of elderly people will be regularly sent and stored in a database.

Furthermore, the system will display all elderly information on a website. Therefore, it will be easy for doctors and relatives to monitor their health status. Since the relatives may be far away from the elderly at the emergency case, the doctors can access the door directly from the website even if it is locked to save time.

The rest of the paper is organized as follows: related work for the proposed system is discussed in Section 2. The proposed system architecture is described in Section 3. The developed database system is presented in Section 4. The system implementation and results are presented in Section 5. Finally, the paper is concluded in Section 6 with a potential for future directions of this work.

II. RELATED WORK

Much work has been conducted to address remote health monitoring and develop systems committed to helping elderly people who live in isolated regions. Wearable devices have contributed to the development of non-invasive strategies to monitor people's activities. Authors in [2] proposed a smart home for elderly care based on a wireless sensor network to monitor and facilitate the elderly. This system integrates many sensors intending to get the required measured parameters: fire detection, gas leakage detection, and the determination of

either door as closed or opened. Another research is proposed in an indoor system based on wireless sensor network technology capable of monitoring heart rate and motion rate. The system is competent to alert whoever cares about the elderly via a smartphone [3]. Other research is proposed to monitor the health of the elderly by measuring the arterial, electrocardiogram, saturation of respiratory oxygen rhythm, blood pressure, and body temperature, then transmits these measurements to a central medical server using Wi-Fi or GSM/GPRS connection [4]. Authors in [5] proposed a system for managing the time of taking medication and monitoring the status based on RFID and wireless sensors.

Moreover, a system that continuously monitors the health state of elderly people and enables them to contact doctors during any emergency cases has been presented [6, 18]. The system can measure human physiological parameters such as breathing frequency, under clothing temperature, heart pulses, positioning inside and outside the house. However, the mentioned systems can monitor the status of elderly people. Some of them still rely on sensors that are not convenient for elderly people.

The industry and research group have conducted projects in health care, for instance, the project called "My-Heart" [11] that focuses on monitoring the measured heart signals to predict and assist patients who are suffering from heart diseases. Further, the project called "HeartCycle" [12, 13] issues a solution for people with Coronary Heart Disease and Heart Failure. Continua Health Alliance has also performed significant enhancement in defining interoperability for both LAN and WAN interfaces and now enables end-to-end interoperability [14, 15, 17].

The proposed system in this paper depends on wearable devices and smartphone devices. The wearable devices are used for tracking data of physical activities. Such tools are used to improve human health by encouraging them to exercise and follow correct behavior in nutrition, sleep, and life. Based on some statistics, about 46% of people older than 65 in the United States own a smartphone [7]. Therefore, this refers to a considerable number of elderly people moving to these new technologies. Many people, especially the elderly, seek a wearable device, which monitors their health and provides essential health parameters to them and their registered doctors [8-10]. Most people are using a wearable watch for fitness, but we use it in this system for taking care of the elderly by integrating a smartwatch with a mobile phone to extract the desired data to analyze them and take actions based on them.

III. PROPOSED SYSTEM ARCHITECTURE

As shown in Fig. 1, the proposed system consists of three units: Elderly, server, and caregivers. Fitbit watch is used to measure the health parameters such as heartbeat, steps, calories, sleep, etc. The smartphone is used as a gateway to send data to the Fitbit servers, and then the data can be retrieved from the Fitbit servers to the local server. Fig. 2 shows the proposed system architecture. Here, we present the main components of the developed platform that targets elderly people monitoring. The proposed platform monitors physiological data of the elderly (e.g., heart rate), as shown in Fig. 1. A wearable device is used, Fitbit is used to collect real-

time data and store it in the Fitbit cloud server via a smartphone. This data will be retrieved using the "OAuth" protocol through API requests.

The developed system comprises a web-based database-driven application that facilitates the sharing and management of useful information about elderly people to an authorized person. The developed system can open the front door of accommodation of older people in case of emergency remotely, provided that caregivers are unreachable. Simply, there is an interfacing action between the doctors and elderly system, see Fig. 1. The block diagram of the proposed system architecture is presented in Fig. 2.

Algorithm 1: System workflow

Output: Elderly status monitoring.

```
do
{
if (a new elderly is registered)
then
{
Assign doctors for the elderly
Register relatives, their locations.
Collect information and update the hospital database server
}
else
{
if (A service is required for the elderly) then
{
Doctors access the website server to monitor the elderly status
if (The status is normal)
Doctor contacts and sends elderly information
else if (The status is critical)
The hospital determines the location of the nearest relative.
Doctors contacts the nearest relative
Send relatives the elderly status
}
If (no relative reply)
The hospital accesses the elderly house directly
}
}
while (true)
```

The used Communication Protocol is known as "OAuth," which is an open-standard authorization protocol or framework that provides applications for "secure designated access [9]. Password data is not sharable if the "OAuth" protocol is used. However, authorization tokens are used to differentiate among both consumers and service providers. "OAuth" protocol is an authentication protocol that allows one to approve one application interacting with another on your behalf without giving away any password. As shown in Fig. 3, it is used for user authorization and API authentication to retrieve the information of the elderly in the Fitbit server to the local server.

The developed website inside the caregiver unit is an interface among sensors readings, doctors, elderly, relatives, and the site map shown in Fig. 4. As shown in Fig. 5, the home page is divided into three sections: login, get started, and

supported platforms. The administrators, elderly, and doctors can use the login section as depicts in Fig. 6 to log in to their pages. The elderly who wants to get started with the developed system can apply a request through the get started section as illustrated in Fig. 6. Then, the elderly will receive an email to confirm that his request has been received, and the administrator of the system will contact him and ask for the required information. The last section is the supported platform, where the people can find the platforms from anywhere, and they can buy products to integrate them with our system.

After the user logs in, he will be redirected automatically to his page. On the admin page, as shown in Fig. 7, there are web site's statistics, which show the number of doctors, the elderly, and the total number of relatives assigned to the existing elderly. Moreover, the admin can view, add, edit, and delete any existing entities on the website. Besides, he can see the elderly list of those who have not been assigned to any doctor and assign them to the proper doctor. Besides, he can reassign any of the elderly to another doctor from one previously

assigned to him. The last section is the messages section. The admin can see the requests from the general people who want to use our system for the elderly. Also, he can see the problem started from the website's users.

The elderly page can be accessed by the elderly himself or by the doctor who has been assigned to him. They can see a live preview of the sensor's readings and graphs, which demonstrate the behavior of each measured parameter. Furthermore, they can generate a full report of elderly health history and print it later to use it wherever needed. Moreover, the doctor can ask the elderly to come to the hospital for a checkup if needed. In case of emergency, he can open/close the door of the elderly house to get the elderly to the hospital if the relatives do not respond to that case, see Fig. 1. On the other hand, the elderly/relative can post questions related to the health status of the elderly to get the answers from the doctor directly via the developed website. Furthermore, there is a record on the elderly page for security issues that shows the name, date, and time of anyone open/close the door from the website.

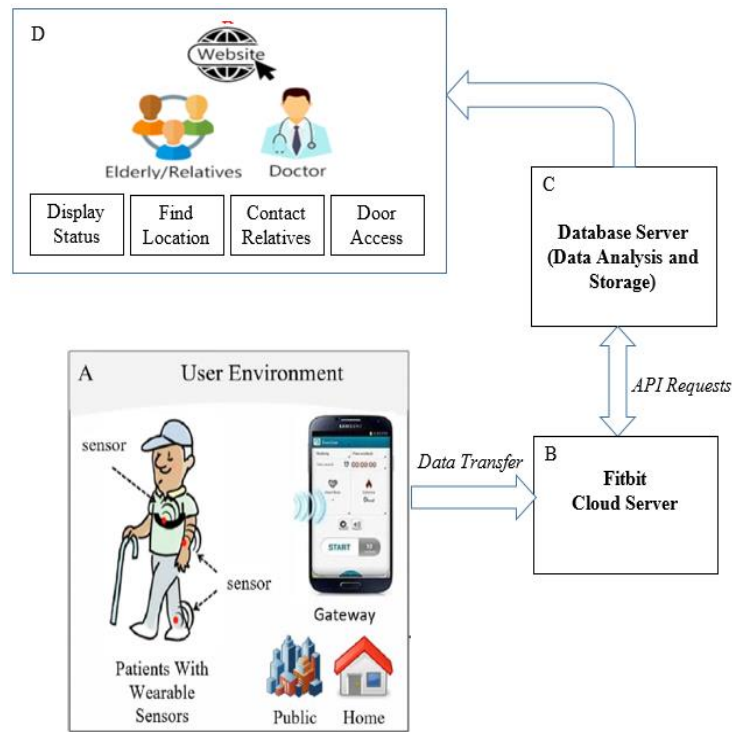


Fig. 1. The Proposed System Overview.

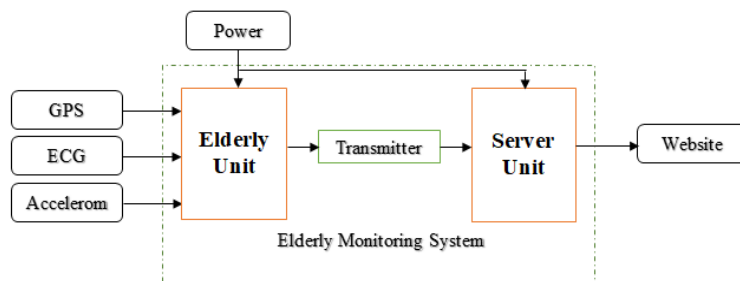


Fig. 2. The Block Diagram of Proposed System Architecture.

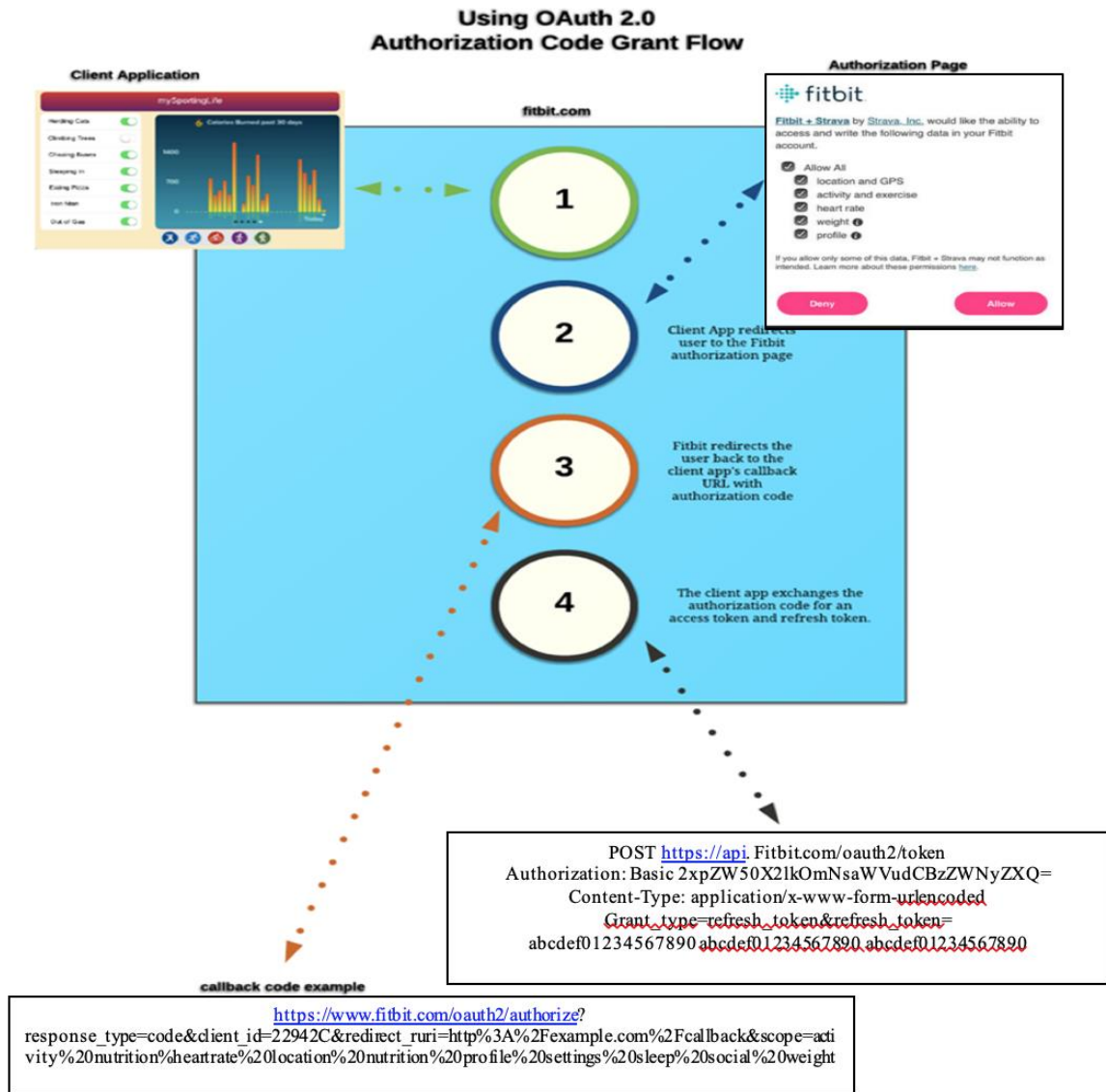


Fig. 3. OAuth 2.0 Protocol in Fitbit [9].

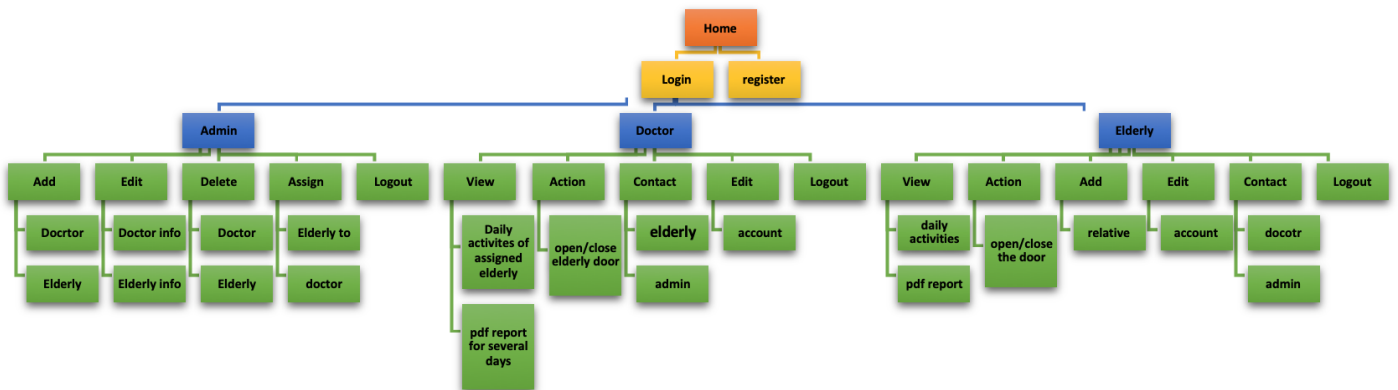


Fig. 4. The Developed Website Map.



Fig. 5. The Developed Home Page.

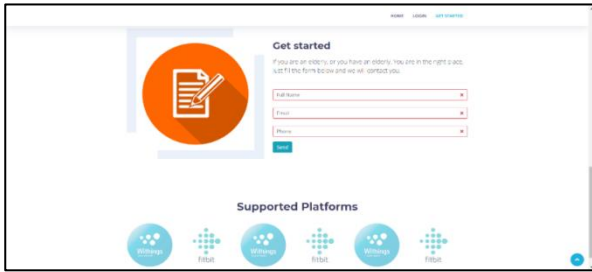


Fig. 6. Login and Support Sections on the Home Page.

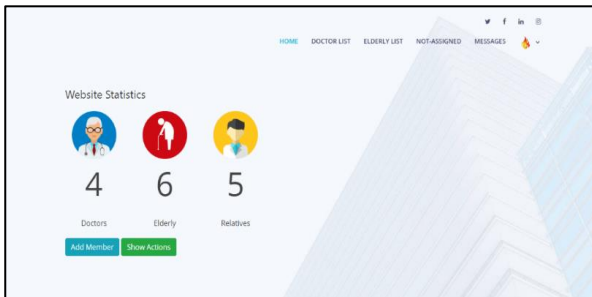


Fig. 7. The Developed Admin Page.

IV. DATABASE MANIPULATION

The database is a collection of organized information so that it can be easily accessed, managed, and updated. Data is represented in the database as tables, and it is indexed to find the relevant information easily. In this study, the database is used to store elderly readings and share them with doctors and relatives through a website [16]. The Entity-Relationship (ER) diagram of the database is shown in Fig. 8.

The business rules describe a policy, procedure, or principle within a specific organization [17-18]. It determines the entities and their relationships on the database. The business rules of the database of the proposed system are presented in Table I.

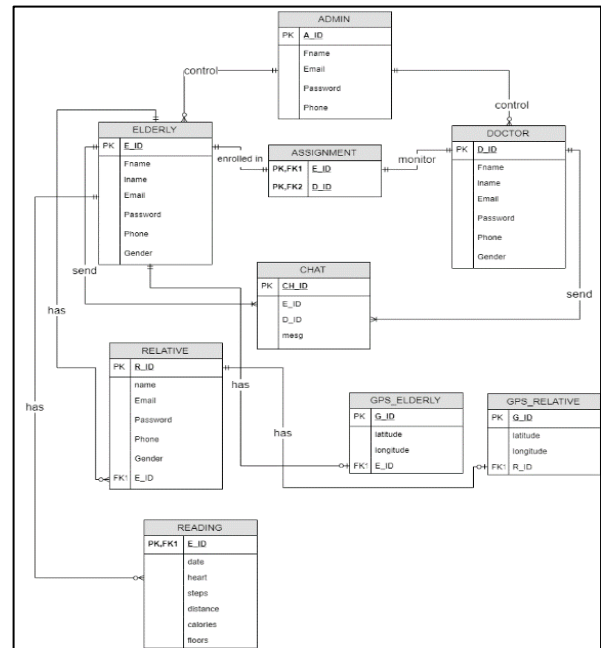


Fig. 8. The Developed ER Diagram.

TABLE I. THE DATABASE BUSINESS RULES OF THE PROPOSED SYSTEM

Steps	Action
1	The elderly have one relative only, but the relative may have many elderly.
2	The doctor may monitor one or more elderly, but each elderly is monitored by one doctor.
3	The elderly store data in history.
4	The Elderly can chat with his doctor.
5	GPS locates the elderly and relatives.
6	Each elderly has a health parameter reading.

V. IMPLEMENTATION

To read the Fitbit watch parameters (heartbeat, steps, calories, sleep, etc.), the Fitbit watch should be connected to the phone using Bluetooth communication. The mobile phone is used as a gateway between the Fitbit watch and the server unit. Fig. 9 shows the data that is read by the Fitbit watch and synchronized with the phone correctly. After uploading the data to the Fitbit server, it should be retrieved and stored in a system database. To retrieve the data, the "OAuth 2.0" protocol (Open Authentication) is used to grant Fitbit and retrieve the required data. This process requires a PHP script that is working continuously. Fig. 10 shows how the script is running and retrieving the data.



Fig. 9. The Connection of Fitbit Watch with the Mobile Phone.

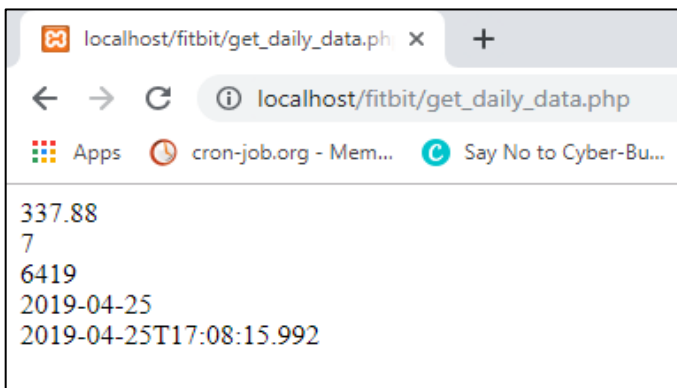


Fig. 10. Get Daily Data Script.

Fitbit provides 150 requests for each user per hour, which forces the designer to manage the requests based on the need for data. For example, daily data will be requested every 3 minutes, so this will cost 50 requests, and the remaining 100 requests will be used to check the connection used with daily activity data, so the remaining is about 50 requests. The left requests will be used to get heart rate data. Elderly and their relatives can contact the doctors through the website; for example, if they need some help or to ask the doctor about an appointment. When an elderly or relative sends a message to the doctor, the message will be saved in the database, and then the message will be sent to the doctor and vice versa (see Fig. 11).

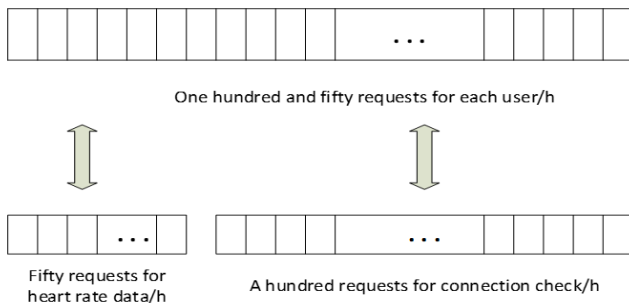


Fig. 11. Requests Provided for users using Fitbit.

Elderly or relatives can contact doctors through the website. For example, if the elderly need some help or needs to ask the doctor about an appointment. The message will be written, saved in the database, and then sent to the doctor and vice versa. The website can add doctors, elderly, and elderly's relatives by the admin that can manage the website. The admin only can add, edit, and remove the user's account, as shown in Fig. 12.

In case of emergency, the medic wants to enter the elderly home, and he needs to control the door of the elderly house. If the door is closed, he needs to open it. By adding a smart lock to the door of the elderly home, doctors can open and close the door remotely from the website. The developed systems support tracking the elderly location if he goes outside the house. The system also calculates, and the doctor can determine the nearest relative to the elderly and sends a notification to him/her.

Fig. 14 shows the location of the elderly, while the blue point and the red points refer to the relatives. Furthermore, a PHP script using Twilio API is implemented to call the relative of the elderly whenever an emergency case occurs. This should work automatically using phone call as shown in Fig. 13.

The integrated system was tested, and the results are shown in Table II. Fig. 12 to 15 shows the system outputs, the recorded heartbeat; Fig. 16 shows the measured data of different activities of the elderly. Fig. 17 gives the statistical data collected during a month.

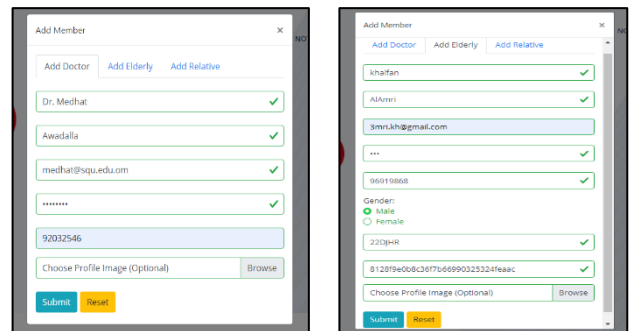


Fig. 12. Adding Doctor and Elderly.

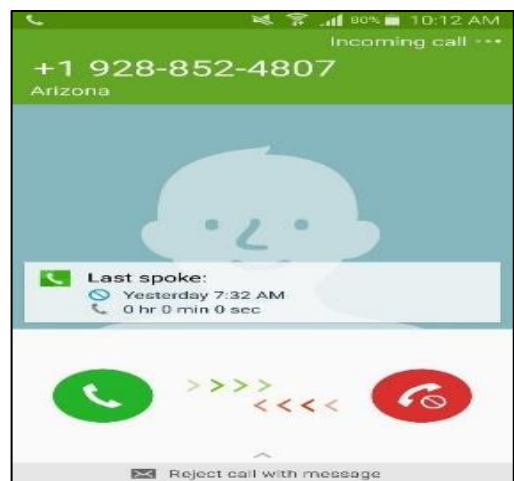


Fig. 13. Track Elderly Location based on GPS Navigator.

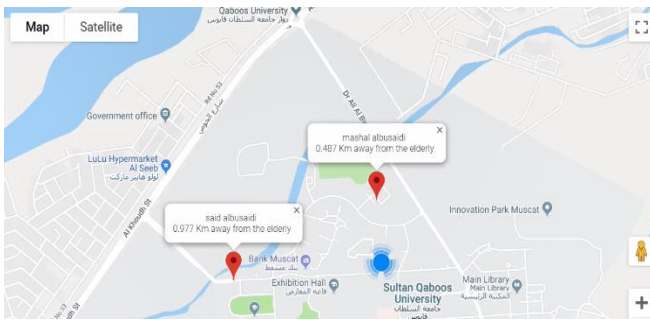


Fig. 14. Contacting the Relative of the Elderly.

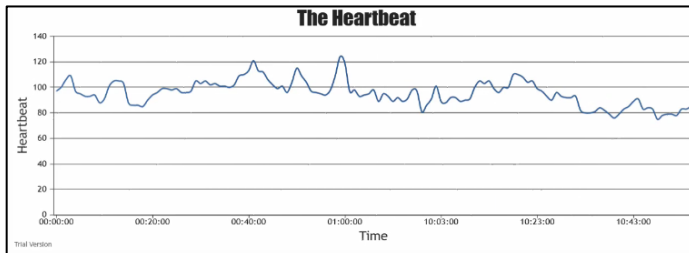


Fig. 15. The Recorded Elderly Heartbeat.

TABLE II. INTEGRATION TEST FOR THE WHOLE SYSTEM

Test name: Integration for the whole system					
Test ID#: IT-01					
Test description: Verify that the project is working					
Test information					
Name of tester: ALL		Date: 25/2/2021			Time: 2:00 PM
#	Procedure	Pass	Fail	N/A	Comments
1	Watch reads all parameters.	√			
2	Watch sends the data.	√			
3	The database receives the data from the watch.	√			
4	The website displays the retrieved data	√			
5	Tracking the live location of the elderly.	√			
6	Measure the distance between the elderly and relatives.	√			
7	Send SMS alarm automatically if there is an emergency.	√			
8	Call relatives automatically if there is an emergency.	√			
9	Open and close the door from the website.	√			



Fig. 16. The Measured Data of the Elderly Activities.

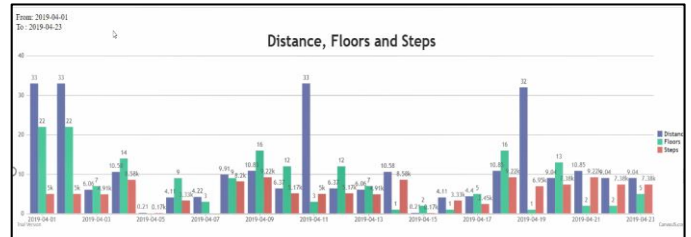


Fig. 17. The Statistical Data of the Elderly.

VI. CONCLUSION

In this paper, an integrated IoT-based platform is presented to monitor and collect vital data from elderly people. The developed platform relies on wearable devices and smartphones. Fitbit is used to collect real-time data and store it in the Fitbit cloud server via a smartphone. OAuth protocol is used to retrieve data stored in the Fitbit cloud server. A web-based database-driven application is developed that facilitates the management of helpful information about elderly people to an authorized person. The stored data is used for analysis, which helps the doctors track their patients' evolution. A complete platform to verify the idea is implemented and tested to validate the system functionality. The results show promising outcomes for monitoring elderly people. Besides, the features of the developed system increase accessibility, efficiency and lower the health expenses that improve the comfortability and safety as well as management of daily routines.

ACKNOWLEDGMENT

The Sultan Qaboos University supports the work in this paper under the internal grant approved for the research project code number IG/ENG/ECED/21/01. Moreover, we are grateful to the anonymous reviewers for their valued feedback.

REFERENCES

- [1] Aging, United Nations, Levels and trends in Population Aging, Available from URL:<http://www.un.org/en/sections/issues-depth/ageing> last update [6/02/2021].
- [2] Rasika S. Ransing, Manita Rajput; Smart Home for Elderly Care, based on Wireless Sensor Network, International Conference on Nascent Technologies in the Engineering Field, 2015, p.p.1-5, 2015.
- [3] Pedro Magaña-Espinoza, Raúl Aquino-Santos, Néstor Cárdenas-Benítez, José Aguilar-Velasco; A Wireless Sensor Network-Based Home Care Monitoring System, IEEE Sensors Journal 12(6):1965-1972, 2012.

- [4] C. Rotariu, H. Costin, Gladiola Andrusac, R. Ciobotariu, F. Adochiei: C. Rotariu, H. Costin, Gladiola Andrusac, R. Ciobotariu, F. Adochiei, "An Integrated System for Wireless Monitoring of Chronic Patients and Elderly People". Proc. of the 15th IEEE Int. Conference on System Theory, Control and Computing, ICSTCC2011, Sinaia, Romania, pp. 527-530, 2011.
- [5] ZigBee Vs. XBee: An Easy-To-Understand Comparison, Available from URL: <https://www.link-labs.com/blog/zigbee-vs-xbee> , Last updated [15/1/2021].
- [6] Ching-Nung Yang, Fu-Heng Wu, Sin-Yen Tsai, Wen-Chun Kuo: E-Health Services for Elderly Care Based on Google Cloud Messaging. IEEE International Conference on Smart City/SocialCom/SustainCom together with DataCom 2015 and SC2 2015, pp. 9-12, 2015.
- [7] Michal Frydrysiak, Lukasz Tesiorowski. Wearable textronic system for protecting elderly people. IEEE International Multidisciplinary Conference on Computer and Energy Science (SpliTech), p.p.1-6, July 2016.
- [8] Mircea Serbanescu, V. M. Placinta, O. E. Hutanu, Cristian Ravariu. Smart, low power, wearable multi-sensor data acquisition system for environmental monitoring. 2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE), 2017.
- [9] Mostafa Haghi, Regina Stoll, Kerstin Thuro. A Low-Cost, Standalone, and Multi-Tasking Watch for Personalized Environmental Monitoring. IEEE Transactions on Biomedical Circuits and Systems, 2018.
- [10] Luprano J, Sol'a J, Dasen S, Koller JM, Ch'etelat O (2016) Combination of body sensor networks and on-body signal processing algorithms: the practical case of MyHeart project. In: null. IEEE, pp 76–79.
- [11] Chouvarda I, Antony R, Torabi A, Weston J, Caffarel J, van Gils M, Cleland J, Maglaveras N (2013) Temporal Variation in telemonitoring data: on the Effect of Medication and Lifestyle Compliance. International Journal of Bioelectromagnetism.
- [12] Palumbo F, Ullberg J, timec A, Furfari F, Karlsson L, Coradeschi S (2014) Sensor network infrastructure for a home care monitoring system. Sensors.
- [13] Wartena F, Muskens J, Schmitt L, PetkovicM(2010) Continua: The Reference Architecture of a Personal Telehealth Ecosystem. In: proceedings of 12th IEEE international conference on e-health networking applications and services (Healthcom), Lyon, France.
- [14] Meredith A. Case, Holland A. Burwick, Kevin G. Volpp, Mitesh S. Patel. Accuracy of Smartphone Applications and Wearable Devices for Tracking Physical Activity Data. JAMA February 10, 2015 Volume 313, Number 6, pp. 625-626, 2015.
- [15] Mobile Fact Sheet, Who owns cellphones and smartphones, Available from URL: <https://www.pewinternet.org/fact-sheet/mobile/> Last updated [10/2/2021].
- [16] C. Coronel, S. Morris, and P. Rob. Database Systems: Design, Implementation, and Management, Boston, Ninth Edition, 2011.
- [17] Zhou, Jining; Zhang, Bo; Tan, Runhua; Tseng, Ming-Lang; Lin, Remen C.-W.; Lim, Ming K. 2020. "Using Neighborhood Rough Set Theory to Address the Smart Elderly Care in Multi-Level Attributes" *Symmetry* 12, no. 2: 297.
- [18] Selvaraj, S., Sundaravaradhan, S. Challenges and opportunities in IoT healthcare systems: a systematic review. *SN Appl. Sci.* 2, 139 (2020)

Study and Analysis for the Choice of Optical Fiber in the Implementation of High-Capacity Backbones in Data Transmission

Wilmer Vergaray-Mendez¹
Facultad de Ciencias e Ingeniería
Universidad de Ciencias y
Humanidades
Lima, Perú

Brian Meneses-Claudio²
Image Processing Research
Laboratory (INTI-Lab)
Universidad de Ciencias y
Humanidades, Lima, Perú

Alexi Delgado³
Interdisciplinary Research Center
Science and Society (CIICS)
Universidad de Ciencias y
Humanidades, Lima, Perú

Abstract—Today, fiber optic implementation projects for backbones have become a necessity, where in many cases failures due to cable stress breaks have been reported. Due to this, it is necessary to carry out a study and analysis of the zone and area prior to implementation. In this research work, through a method based on theory and analysis, the geographical and climatological conditions where the optical fiber will be installed in the Lima region, Peru, will be evaluated, as well as the study of mechanical loads and electric fields associated with the installation of fiber optics on existing electrical network lines will also be carried out. The results of this study showed that for regional backbone projects in the city of Lima, Peru, the use of the type of optical fiber should be considered under the recommendation of the International Telecommunications Union (ITU) -T G.652.D All -Dielectric Self-Supporting (ADSS). The studies and results obtained in this research may also help the various companies in the sector, in future implementations of high-capacity fiber optic backbones in data transmission, to make the best decision on the type of cable and its recommended characteristics for the region.

Keywords—Mechanical loads; electric fields; backbones; optical fiber; data transmission

I. INTRODUCTION

Optical fibers constitute the central axis of the global telecommunications system, they were designed to transmit high information capacity [1]. The growing demand in speed of information transmission has made the implementation of fiber optics comply with national and international standards and recommendations [2]. The absence of orderly information and poor practices followed by empiricism in the choice and installation of fiber optics in the Lima-Peru region, have led to high numbers of expenses for operational maintenance of the network, as shown in Table I, because it is not considered the geographical and climatological factors and conditions of the Lima region, based on the National Electricity Code (CNE) [3]. That is why in the present research work, based on the CNE and considering the factors and conditions mentioned above, a study and analysis was carried out to choose the type of fiber optic suitable for installation in the Lima-Peru region.

Table I shows a comparative table in percentage of what the operators represent to carry out the maintenance of the fiber optic network with respect to the annual capex in Peru.

The optical fiber is resistant to deterioration and high temperatures, the protection of the coating is ideal to withstand high stresses in the installation, the optical fiber does not originate electromagnetic radiation, it is resistant to intrusive actions, to access the information that circulates in the fiber it is necessary to cut it, which there is no transmission of information during this process, and it is easily detectable, therefore it is a safe means of transmitting information [4]. However, a common problem that usually occurs when choosing the type of fiber optic cable is that an exhaustive analysis of the mechanical loads to which the cable will be exposed in the installation is not carried out, thus verifying the maximum allowable stresses of the fiber optic cable ADSS [5]. Due to this, the present research work will analyze the mechanical loads that affect the optical fiber when it is installed in the Lima-Peru region.

Choosing a fiber optic cable for any application implies considering in an analytical way the electric field levels produced by medium voltage lines in which they are installed to allow the network to be expanded throughout the entire national and international territory [6]. For the installation, it is necessary to consider where and how the cable will be installed, either in an underground conduit in an outside plant or in trays within a building. Requirements for long-term installations include humidity or exposure to water, temperature, tension in overhead cables, and environmental factors in the various areas of the region Perú [7].

TABLE I. COMPARATIVE FIBER OPTIC NETWORK MAINTENANCE

Year	G.652.D	Others
2020	6% of Annual Capex	14% of Annual Capex
2019	7% of Annual Capex	16% of Annual Capex
2018	9% of Annual Capex	19% of Annual Capex

In the Thesis Design of a Broadband Network in the San Martin Region, Peru, Carlos Bedregal, indicates that fiber optic technology is an excellent means of transporting information, even more so through new scientific developments in the field of optics, which give it a high scalability factor in the present [8]. That is why the thesis presents as a highly effective and profitable alternative the implementation of fiber optics on existing electrical network infrastructure, to provide telecommunications services to many users, in addition to being able to comfortably support the increase of users in the future, without having to make new capital investments.

In the Thesis Design of a Fiber Optic Network for Broadband in Coishco, Ancash, Peru, Elliot Lopez, recommends the use of fiber optic cable that complies with the ITU-T G.652.D [9], due to the fact that long distance communication systems are optimal for solutions, a main characteristic is the high transmission capacity due to the increase in bandwidth, it allows reaching up to 3000 km and 40 km of distance at speeds of 10 Gbps and 40 Gbps respectively without regenerations, it is also ADSS type, essential for installation on infrastructure of medium and low voltage power lines. [10].

In [5], it mentions that the fiber optic ITU-T G.652.D presents an optimal performance in the operation of the transmission equipment of the highest range Dense Wavelength Division Multiplexing (DWDM) according to manufacturer's NEC, Huawei, Alcatel Lucent and other fiber optic providers, likewise, it does not present disadvantages compared to other types of optical fibers such as G.655 and G.656. Optical fiber under ITU-T Recommendation G.652.D is designed for high-capacity backbone networks and operates with DWDM technology from leading DWDM equipment manufacturers.

The increasing requirement for capacity and speed in the transmission of information has made fiber optics the preferred transmission medium in the construction of physical communication links [2], that is why the objective of this research article is oriented in the study and analysis of the various factors, such as geographical and climatological conditions that the Lima region presents, in order to choose the type of optical fiber with specific characteristics for certain areas of the Lima region, Peru, to achieve this a series of hypotheses is considered for the various areas that the region presents, all this analysis is carried out based on the standards and norms that include the design, manufacture, installation and testing of fiber optics [11]. The Lima-Peru region is specifically considered for its geographical areas that are in rugged areas, difficult to access, extreme and variable climates as shown in Fig. 2, where the analysis of mechanical loads and electric fields can be carried out, which are considered as a fundamental part in the research work, since transmitting high information capacity requires that the physical environment be as optimal as possible.

The study and analysis of the geographical and environmental conditions were analyzed with the help of the National Electricity Code, which describes the characteristics of the various zones and load areas of the Lima region, Peru, after that to acquire the values of electric field produced by

medium and low voltage lines, the calculation software Electric Field of Transmission Lines (EFT), an application of the Electric Power Research Institute (EPRI), was used.

This research work is structured as follows: In Section II, the theoretical description, types, and characteristics of the fiber optic cable will be shown, as well as the technical standards associated with it. In Section III, the study and analysis for the choice of optical fiber will be shown, where the mechanical loads and electric fields to which the optical fiber will be exposed at the time of installation are described. In Section IV, the results obtained when carrying out the analysis and study of the calculations of mechanical loads and electric fields will be shown. In Section V, the discussions of the research work are presented and finally in Section VI, the conclusions, as well as the recommendations of the future research that will be achieved with this article.

II. OPTICAL FIBER

A. Types of Fiber Optics

According to the number of modes of propagation of the light ray, an optical fiber is classified into single-mode and multimode, both of which have a diameter of 12 μm coating. The ITU defines six operating bands for single-mode optical fiber, according to wavelength (λ): Original (O), Extended (E), Short wavelength (S), Conventional (C), Long wavelength (L), Ultra wavelength (U), as shown in Fig. 1 [5].

1) *Multimode optical fibers*: Multimode optical fibers, on the other hand, have a diameter of 9 μm . They can transmit several light rays by successive reflections, this being their main limitation for their consideration in the design of backbones networks [5].

2) *Single-mode optical fibers*: Single-mode optical fibers have a core that varies from 7 to 11 μm , due to their special design they have the ability to send a single mode of light, understood as "light mode" to a "light ray" in theory only allows the propagation of a straight-line light mode using a high intensity laser cannon, furthermore, they allow reaching long distances and high data bit rates [5], as shown in Table II.

Table II shows the OM* nomenclature according to the Telecommunications Industry Association (TIA), B* nomenclature according to the International Electrotechnical Commission (IEC), G* nomenclature according to the International Telecommunications Union (ITU) [12].

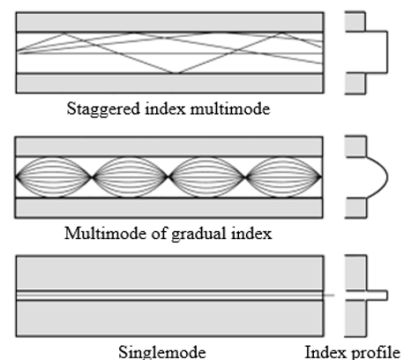


Fig. 1. Types of Optical Fiber.

TABLE II. TYPES OF OPTICAL FIBER AND SPECIFICATIONS

Core/Coating	Attenuation	Bandwidth	Applications/Notes
Multimode	850/1300nm	850/1300nm	
50/125 microns (OM2)	3/1 dB/km	500/500 MHz-km	For lasers for GbE LAN networks
50/125 microns (OM3)	3/1 dB/km	2000/500 MHz-km	Optimized for 850 nm VCSEL
50/125 microns (OM4)	3/1 dB/km	4700/500 MHz-km	Optimized for 850 nm VCSEL>10Gb/s
62.5/125 microns (OM1)	3/1 dB/km	160-200/500 MHz-km	Fiber for LAN (FDDI)
100/140 microns	3/1 dB/km	150/300 MHz-km	Obsolete
Single mode	1310/1550nm	1310/1550nm	
9/125 microns (OS1, B1.1, or G.652)	0.4/0.25 dB/km	~100 Tz	Telecommunications / Cable TV, long distance, and high-speed LAN networks
9/125 microns (OS2, B1.3, or G.652)	0.4/0.25 dB/km	~100 Tz	"Low Water Peak" (LWP) Fiber
9/125 microns (B2, or G.653)	0.4/0.25 dB/km	~100 Tz	Dispersion Shifted Fiber (DSF)
9/125 microns (B1.2, or G.654)	0.4/0.25 dB/km	~100 Tz	Cut-off Shifted Fiber (CSF)
9/125 microns (B4, or G.654)	0.4/0.25 dB/km	~100 Tz	Non-zero dispersion-shifted fiber (NZ-DSF)

3) Technical standards

- Código Nacional de Electricidad, Perú (CNE) Reglas 250.B, 250-C, Tabla 235-5 [3].
- Institute of Electrical and Electronics Engineers (IEEE) Std 1222-2011, Standard for Testing and performance for ADSS Fiber Optic Cable for Use on Electric Utility Power Lines - Std 776™-1992 (R2008) IEEE Recommended Practice for Inductive Coordination of Electric Supply and Communication Lines - Std 1137™-1991 (R2008) IEEE Guide for the Implementation of Inductive Coordination Mitigation Techniques and Application [13].
- Resolución Ministerial-368-2011-MTC / 03, Working Document "Especificaciones Técnicas para el Tendido de Fibra Óptica en Redes de Energía Eléctrica y de Hidrocarburo" [14], [15].
- Single-mode ADSS fiber optic cable data sheets, span 250, 600, 1000 and 1200 [16].

III. METHODOLOGY

A. Recommendation ITU-T G.652

It is a standard single-mode optical fiber with non-shifted dispersion, it was initially optimized for use in the 1310nm

wavelength region, but it can also be used in the 1550nm region, it presents "zero" dispersion (0.092 ps / nm² * km) in the second window at 1300nm to 1324nm, it is suitable for Coarse Wavelength Division Multiplexing (CWDM), it can be used in the second window with worse attenuation at 0.4 dB/km or in the third window with worse dispersion, but can be improved with compensating devices dispersion [9]. The ITU-T G.652 recommendation has four variants A, B, C and D. The G.652 A and B optical fibers were the first to be marketed, but they were discontinued because they presented high attenuation in the 1390nm band (E band), phenomenon known as "water peak", which was corrected in later versions G.652 C and D, with this optimization the E band can be fully exploited.

In this research, emphasis is placed on the use of fiber optic cable manufactured based on the recommendations of ITU-T G.652.D, because its characteristics are the best option for installation in geographic areas of the Lima region; below in Table III the most outstanding is described:

TABLE III. CHARACTERISTICS OF THE ITU-T G652.D FIBER OPTIC

Description	G.652.D	
Modal Field Diameter (mm)	1310 nm	9.2 ± 0.4
	1550 nm	10.3 ± 0.5
Attenuation Coefficient (dB/Km)	1310 nm	≤ 0.35
	1383 nm	≤ 0.35
	1550 nm	≤ 0.24
Chromatic Dispersion (ps/nm. Km)	1285 – 1330 nm	≤ 3
	1550 nm	< 18
Zero Dispersion Wavelength (nm)	1300 – 1322	
Zero Dispersion Slope (ps / nm ² Km)	≤ 0.092	
Refractive Index	1310 nm	1.467
	1550 nm	1.468
Cut Wavelength (nm)	Cabling	≤ 1260
PMD (ps / (ps/√Km) Link Value)	1550 nm	≤ 0.1
Mechanical Specifications	G.652.D	
Core non-circularity	≤ 6 %	
Core / cladding concentricity error	≤ 1 mm	
Cladding diameter	125 ± 1 mm	
Cladding non-circularity	≤ 1 %	
Primary coating diameter	245 ± 10 mm	
Primary coating non-circularity	≤ 6 %	
Primary coating concentricity error	≤ 12.5 mm	
Proof Test	≥ 8.8 N / ≥ 1 % / ≥ 100 Kpsi	

B. Study of Mechanical Loads of Fiber Optic Cable

The study and analysis to be carried out in this article is based on the CNE, where we will carry out an exhaustive analysis with different hypotheses considering the best and worst geographical, environmental and climatological conditions for the choice of the type of Optical fiber to be used in the implementation of a backbone network for transmission of high capacity information, those are detailed in Table VI and

VII [17], which is in the order superior to Gb/s. According to the CNE, the Lima region is located in Zone B, Loading Area 0, 1, 2 and 3, as shown in Fig. 2; which for the purposes of the study will only take specific data and has been classified by Cluster 1 and 2 (North) and Cluster 3 and 4 (South) [3], as described in Tables IV and V.



Fig. 2. Location of Load Areas in Peru.

The purpose of this calculation is to verify the maximum admissible stresses of the fiber optic cable ADSS span 250, 600 and 1200, for its subsequent implementation in existing and projected electrical networks, which is part of this study article [5] The calculations will be modeled taking into consideration the recommendations given in the CNE [3], and the techniques of FiberHome, manufacturer of the fiber optic cable [16].

1) Geographic location of cluster 1 and 2

TABLE IV. FIBER OPTIC SECTIONS

SECTION A-B	PROVINCE	START DISTRICT	FINAL DISTRICT	CLUSTER
SL09 - COPA	CAJATA MBO	HUNCAPON	COPA	1
OYON - CHURIN	OYON	OYON	OYON	1
SAYAN - HUMAYA	HUAURA	HUAURA	HUAURA	2
HUAURA - HUMAYA	HUAURA	HUAURA	HUAURA	2
ACOS - CARAC	HUARAL	ACOS	CARAC	2

2) Climatic characteristics: The climatological characteristics correspond to the study areas as stipulated in the CNE Section 25, Rules 250B and Table 250-1 [3].

TABLE V. LOCATION ACCORDING TO ZONE AND AREA

Section	Zone	Area	Long	Start	Alt	End	Alt
SL09 - Copa	B	1,2	16635	1	3563	92	3368
Oyon - Churin	B	0,1	25655	1	3702	236	2284
Sayan - Humaya	B	0	25309	1	637	330	390
Huaura - Humaya	B	0	23296	1	64	266	321
Acos - Carac	B	0	11394	1	1853	37	2466

3) Characteristics of the existing power line: For study and calculation purposes, the fiber optic cable to be analyzed has the projection to be implemented over the existing electrical network, considering a length of approximately 750 km, being the following for each cluster: Cluster 1: 265 km and Cluster 2: 485 km. The following main characteristics are considered:

- Nominal voltage: 10, 13.2, 22.9, 60 kV.
- Maximum voltage: 11, 14.5, 25, 66 kV.
- Electric Company: Adinelsa, Enel, Coelvisac.
- Number of Circuits: Simple and double.
- Conductor arrangement: Vertical, horizontal, and triangular.
- Length: 510 km.
- Active Conductor: 01 conductor per phase All Aluminum Alloy Conductors (AAAC) 35-120 mm².
- Structures: Concrete, metallic, fiber and wood posts.
- Types of Arming: Suspension, anchoring, and retention.

4) Mechanical calculation of the fiber optic cable: The mechanical calculations of the cable aim to determine the following relative magnitudes in all the analysis hypotheses:

- Horizontal force of the conductor.
- Tangential force of the conductor in the supports.
- Driver's arrow.
- Driver parameters.
- Exit angles of the conductor with respect to the horizontal line, in the supports.
- Span - weight of the structures.
- Span - middle of the structure.

5) Maximum stresses on the conductor

a) Cable stress in Everyday Stress (EDS): The technical standards regarding the behavior of conductors recommend using the following horizontal effort as a reference:

- In the Everyday stress (EDS) condition 18% of the cable breaking stress according to good industry practices.

b) Maximum cable stresses: The maximum stresses in the cable are the tangential stresses that occur at the highest points of the catenary.

- For ADSS fiber optic cable, its value should not exceed 51% of the breaking stress. (Value given by the fiber optic cable manufacturer).

6) Calculation hypothesis: The state hypotheses for the mechanical calculations of the fiber optic cable are defined on the basis of the following factors, the same as those defined by the CNE [3], and the following hypotheses will be considered: Wind speed, temperature and ice load.

TABLE VI. ZONING, AREAS, AND LOADS

Hypothesis 1	Longer Duration Condition (EDS)
Temperature	Average annual
Wind speed	Null
Ice overload	Null
Hypothesis 2	Minimum temperature condition
Temperature	Minimum
Wind speed	Null
Ice overload	Null
Hypothesis 3	Maximum wind speed condition
Temperature	Minimum
Wind speed	Maximum
Ice overload	Null
Hypothesis 4	Maximum ice load condition
Temperature	Medium
Wind speed	Null
Ice overload	Maximum
Hypothesis 5	Maximum temperature condition
Temperature	Minimum
Wind speed	Null
Ice overload	Null
Hypothesis 6	Wind and ice combine
Temperature	Minimum
Wind speed	Maximum
Ice overload	Maximum

TABLE VII. MAXIMUM VALUES OF HYPOTHESES

	Area 0	Area 1	Area 2	Area 3
	< 3000 masl	3000 a 4000 masl	4001 a 4500 masl	>4500 masl
Hypothesis 1	Unit of Measure °C			
Longer lasting condition - EDS				
Average annual temperature	15	10	5	0
Wind speed	0	0	0	0
Ice overload	0	0	0	0
EDS (% Breaking pull)	18	18	18	18
Hypothesis 2	Unit of Measure °C			
Minimum temperature				
Minimum temperature	5	0	0	-10
Wind speed	0	0	113	0
Ice overload	0	0	0	0
% Breaking pull - ADSS	51	51	51	51
Hypothesis 3	Unit of Measure °C			
Maximum wind speed				
Average annual temperature	10	5	0	-5
Wind speed	94	104	113	120
Ice overload	0	0	0	0
% Breaking pull - ADSS	51	51	51	51
Hypothesis 4	Unit of Measure °C			
Maximum ice load				
Minimum temperature	0	0	-5	-10
Wind speed	0	0	0	0
Ice overload	0	6	25	50
% Breaking pull - ADSS	51	51	51	51
Hypothesis 5	Unit of Measure °C			
Maximum temperature				
Maximum temperature-ADSS	30	25	20	15
Wind speed	0	0	0	0
Ice overload	0	0	0	0
% Breaking pull - ADSS	51	51	51	51
Hypothesis 6	Unit of Measure °C			
Wind + ice				
Average annual temperature	5	0	-5	-10
Wind speed	50	52	56	61
Ice overload	0	3	12	25
% Breaking pull - ADSS	51	51	51	51

7) Formulas considered

a) Equation of Change of State: The change of state of the cables for different spans and different environmental conditions is governed by the following cubic equation [18].

$$T_f^3 + T_f^2 * \left[\frac{d^2 W_i^2 E \cos^3 \phi}{24 S \sigma_i} + \alpha (t_2 - t_1) E \cos \phi - \sigma_i S \right] - \frac{d^2 W_f^2 E \cos^3 \phi}{24} = 0$$

$$\cos \phi = \frac{1}{\sqrt{1 + \left(\frac{H}{D}\right)^2}} \quad (17)$$

Where:

- T_f : Final horizontal draft (kg)
- d : Vain (m)
- W_i : Initial unit weight (kg/m)
- W_f : Final unit weight (kg/m)
- t_2 : Conductor section (mm²)
- σ_i : Initial unit horizontal force (kg/mm²)
- t_2 : Final temperature (°C)
- t_1 : Initial temperature (°C)
- α : Coefficient of linear expansion (1/°C)
- E : Elastic modulus (kg/mm²)
- $\frac{H}{D}$: = Ratio of unevenness/span

With the equation, the results of the mechanical calculation of the various environmental conditions that affect the fiber optic cable are obtained, they will be analyzed in zone B, Areas 0, 1, 2 and 3 for the different state hypotheses previously described.

8) Wind pressure: The wind pressure has been determined according to what is indicated in the CNE, which indicates in rule 250.B and table 250-1-A, that the speed should be 94 km/h for load zone B (moderate). To obtain the wind pressure, the formula of CNE rule 250.C is applied [3].

$$P_v = K * V^2 * S_f * A \dots [3]$$

Where:

- P_v : Load in Newton
- K : Pressure Constant
- K : 0.613 for altitude up to 3000 masl.
- K : 0.455 for altitude greater than 3000 masl.
- V : Wind speed m/s
- S_f : Form factor (see Rules 251.A.2 y 252.B.2)
- A : Projected area m²

With the equation, the results of the calculations of the horizontal wind loads or pressures due to the wind that are applied to the fiber optic cable are obtained in the areas where it is planned to install, according to the different hypotheses previously raised.

9) Distribution of structures with fiber optics: The following in the Fig. 3 graphically shows the typical distribution of ADSS fiber optic cable, located under the phase conductors of existing electrical networks.

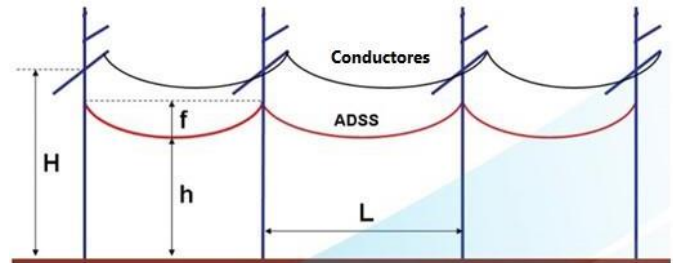


Fig. 3. Line of mean Voltage Line.

Where:

- H : Height of electrical structure
- f : fiber optic arrowing
- H : Height of the optical fiber with reference to the ground
- L : Span length
- ADSS : Fiber Optic Cable
- Conductors: Power line cable

10) Calculation methodology

a) Data entry: The analysis begins with the entry of information from the topographic survey and technical characteristics of the existing electrical network given in the data portfolio. The technical characteristics of the fiber optic cable of each span 250, 600 and 1200 are considered, such as: section, diameter, weight of the cable, breaking pull, elasticity modulus and coefficient of expansion [16]. Six state hypotheses are established that allow the simulation of the mechanical calculation of the cable in the most unfavorable scenario through the State Change Equation (ECE) [18].

b) Data reporting: The calculations of mechanical stresses of the fiber optic cable, analyzed in each state hypothesis, are shown in a report sheet. It is verified that in the most unfavorable scenario it does not exceed the maximum allowable tension defined by the manufacturer ($\leq 51\%$). Likewise, in each section analyzed, the gap with the maximum arrow is identified; for subsequent verification of compliance with the Minimum Safety Distance (MSD).

C. Technical Studies of Electric Fields

The purpose of the development of this study is to determine in an analytical way, the electric field levels produced by medium voltage lines, which are distributed in 47 districts within the Lima Region, the voltage levels presented

by these power lines correspond to 10kV and 2.9 kV. The calculations to be carried out will have the purpose of verifying compliance with the levels of exposure to magnetic fields allowed by the CNE [3], the International Commission on Non-Ionizing Radiation Protection (ICNIRP) [19], and IEEE Standards Association, "IEEE 1222-2019 - IEEE Standard for Testing and Performance for ADSS Fiber Optic Cable for Use on Electric Utility Power Lines," Nov. 07, 2019 [13].

The electric power transmission line produces electric field emissions, which vary according to the characteristics and disposition of the conductors [19]. These emissions must respect the maximum permissible exposure limit values established in National and International Regulations.

For the case of this study, the analysis of electric fields was carried out at the installation height of the fiber optic cable, in a transverse direction to the axis of the line up to the limit of the easement strip, as established in Section 21, rule 212 of the CNE.

As indicated in Resolución Ministerial N° 368-2011-MTC/03 "Especificaciones Técnicas para el Tendido de Fibra Óptica en Redes de Energía Eléctrica y de Hidrocarburo"; section 2.4; rule 2.4.1 [15]. For aerial applications, in the case of ADSS [13], the following should be considered:

- Strong electric fields, overhead cables without metal parts installed in the high-voltage and medium-voltage environment of power transmission lines are subject to the influence of the electric field of these power lines, which can lead to phenomena such as corona effect, arcing or conducting path in cable cover [20].
- To avoid damage, the fiber optic cable must be installed in electric transmission lines considering values where the magnetic electric field is minimal, likewise it is important to use special cable covering materials depending on the level of electric/magnetic field present in the work area.

11) *Characteristics of the medium voltage line:* For the analysis of medium voltage lines, the approximate distances of the existing electrical networks in the project area are presented, which are distributed in 325 km of electrical network belonging to the ENEL company, 10.24 km of electrical network belonging to the company COELVISAC and 207.76 km of electrical network belonging to the company ADINELSA.

a) *Installation Distance of Fiber Optic Cable:* The vertical safety distance between conductors and communication cables should not be less than indicated in table 233-1 of the CNE [3]. For the installation of the fiber optic cable, a minimum vertical distance of 1.80m in medium voltage line and 0.60m in low voltage line must be considered. To comply with the provisions of the CNE, section 23, a safety distance of 2.30m has been considered between the phase conductor and the fiber optic cable, in the same way as shown below in Fig. 4.

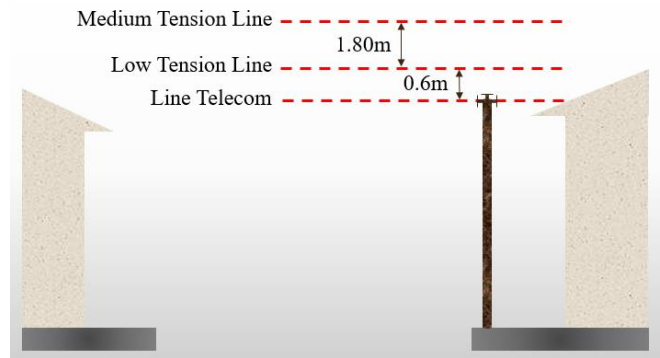


Fig. 4. Safety Distances from Installation of Medium and Low Voltage Networks to Communications Networks.

12) *Electric field analysis:* The electric field describes the force exerted on a unit of electric charge. It is represented by lines of force that surround the electrical device. Electric fields are produced due to the presence of electric charges, regardless of their state of motion [21]. A charge at a given point produces an electric field in all directions, with a spherical symmetry pattern. In an electric transmission line, it produces an electric field around its entire path, giving rise to a model with cylindrical symmetry, in the same way as shown below in Fig. 5 [6].

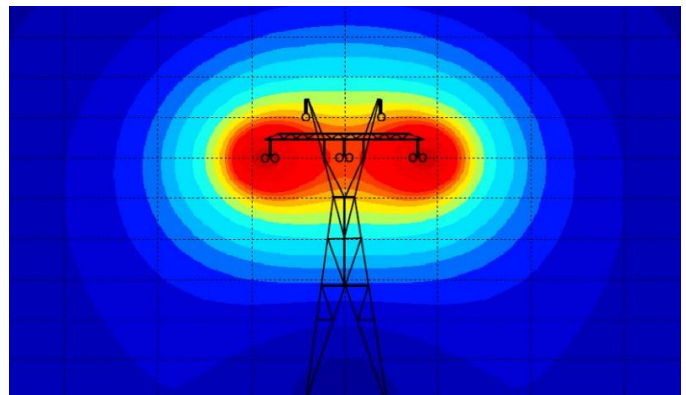


Fig. 5. Electric Field Generated by Electric Cables.

a) *Tracking effect:* The use of existing structures of electrical power networks for the installation of fiber optic cables is quite common, which generates significant savings compared to other forms of installation [15]. However, although the cost-benefit ratio is quite interesting, this methodology also involves some difficulties. A quite common problem is the erosion of the outer covering of the optical cables caused by the tracking effect, due to the proximity with the high voltage cables and the existing electric field [6]. Explaining better: the exposure of the cable to the electrical potential generated by the power transmission line, added to the effect of bad weather (rain, solar radiation, etc.) and environmental pollution, results in the emergence of conductive regions on the surface optical cables.

IV. RESULTS

Under these conditions, when two humid regions delimit an intermediate dry region, the emergence of a potential gradient occurs. The higher the level of pollution in the environment, the more intense the electric discharge generated between the humid regions, since the pollution particles deposited on the surface of the cable increase the level of electric current now of discharge. The small and constant electrical discharges that arise between the two humid regions, finally, cause the heating, the breaking of the polymeric chains and other chemical reactions in the material that constitutes the outer cover of the optical cables. This effect is accompanied by light scintillations (leakage currents) and ends up forming a "trace", which is a permanent conductive path that appears on the surface of the insulating material. This is the phenomenon of the "Tracking Effect" also known as "Dry Band Arcing" [22], which ends up damaging the material of the outer cover of the optical cable, totally compromising its protection, as can be seen in Fig. 6 [23].

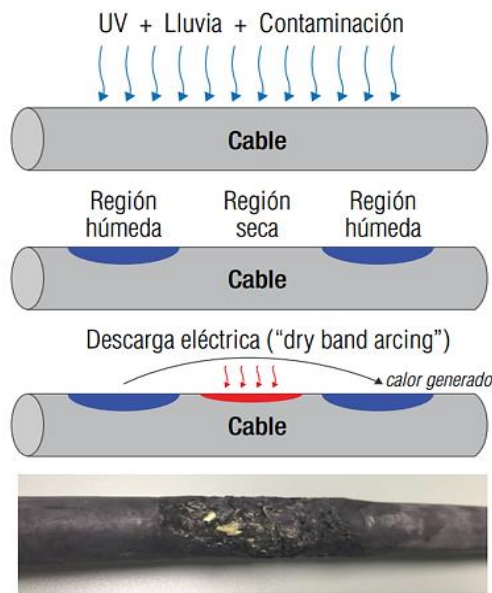


Fig. 6. Optical Cable Degraded by the Tracking Effect.

In accordance with Standard IEEE 1222-2019, the way to verify the tracking effect and the resistance of its cover to erosion are indicated in the tests on ADSS cables, for cables with Class A cover, no electrical tests are required. A cable with a Class A cover, it is understood that it will be used when the high voltage line is below 110kV and therefore the level of electrical stress will not exceed 12kV of space potential acting on the cable cover [13]. For ADSS cables with Class B cover, it implies that the cable may be used in high voltage lines above 110kV and therefore require electrical tests, it corresponds to a level of electrical stress that exceeds 12kV of space potential acting on the cable cover.

13) Calculation methodology: The electric field calculations were developed using the Electric Field of Transmission Lines (EFT) calculation software, an application of the Electric Power Research Institute (EPRI) [24].

This section will show the results of the studies carried out, such as the mechanical loads and electric fields to which the fiber optic cables will be exposed when installed on existing electrical network lines.

A. Result of the Study of Mechanical Loads

The results obtained in the mechanical calculation of the fiber optic cable, the same ones that were analyzed in zone B, areas 0, 1, 2 and 3 for the different state hypotheses, as shown in the following report, where it is concluded that the fiber optic cable to be installed is below the maximum allowable tension given by the manufacturer, so its installation is feasible. The results of the modeling of the mechanical calculation of ADSS fiber optic cable are shown in the following Tables VIII and IX, which for calculation purposes only the most unfavorable scenarios have been considered.

TABLE VIII. MECHANICAL LOADS OF THE FIBER OPTIC CABLE CARGAS MECÁNICAS DEL CABLE DE FIBRA OPTICA

Study of Mechanical Loads of Fiber Optic Cable	Ring	C1	C1	C2
	Section	S109 - copa	Oyon - Churin	Sayan - Humaya
	Item	9-10	154-185	92-93
	East	273157	295260	255096
	North	8839097	8806820	8767631
	Altitude (masl)	2823.00	2379.00	586.48
	Span	600	250	1200
Span	a(m)	390.39	213.86	84.01
Unevenness	h(m)	-141.00	-2.00	-0.82
Area	Area	AREA0	AREA0	AREA0
Hypothesis 1	Shot H(kg)	407.97	229.74	773.56
	TMax (kg)	444.65	230.33	773.70
	Arrow (m)	7.00	3.31	0.20
Hypothesis 2	Shot H(kg)	409.91	232.38	774.86
	TMax (kg)	446.72	232.96	775.00
	Arrow (m)	6.97	3.27	0.20
Hypothesis 3	Shot H(kg)	780.23	439.69	815.02
	TMax (kg)	881.48	444.86	815.83
	Arrow (m)	15.76	7.71	0.73
Hypothesis 4	Shot H(kg)	410.89	233.71	775.51
	TMax (kg)	447.75	234.29	775.65
	Arrow (m)	6.95	3.25	0.20
Hypothesis 5	Shot H(kg)	405.07	225.84	771.62
	TMax (kg)	441.58	226.43	771.76
	Arrow (m)	7.05	3.37	0.20
Hypothesis 6	Shot H(kg)	409.91	232.38	774.86
	TMax (kg)	453.50	233.69	775.07
	Arrow (m)	10.80	5.18	0.29

TABLE IX. MAXIMUM STRESS OF THE FIBER OPTIC CABLE

	Section	Oyon - Churin	SI09 - Copa	Sayan - Humaya
	Cluster	1	1	2
Maximum Effort (kg) - Calculated	Span 250	444.86	899.67	408.36
	Span 600	916.47	1671.76	1487.31
	Span 1200	1,663.40		
Breaking Load (kg) - Established by the Manufacturer	Span 250	1,276.36	1,276.36	1,276.36
	Span 600	2,266.50	2,266.50	2,266.50
	Span 1200	4,297.57	4,297.57	4,297.57
Verification% of Maximum Span Effort	Span 250	35%	40%	32%
	Span 600	40%	39%	35%
	Span 1200	0.39		
Permitted →		≥ 51%	≥ 51%	≥ 51%

B. Result of the Study of Electric Fields

From the results obtained, for the conductor arrangements of the 10 and 22.9kV medium voltage lines, it is concluded that the ADSS type fiber optic cable to be installed complies with the provisions of the IEEE 1222-2019 standard, title 5 ADSS Application Requirements and recommendations”, item 5.9 “Electrical Requirements (electric fields, corona effect and contamination)”, as well as the provisions of the CNE, section 21, rule 212 "Induced Voltages - Electric Fields"[1], by the following considerations:

1) The ADSS fiber optic cable to be installed in the most unfavorable scenarios is immersed in an electric field lower than that established in the CNE (< 8.3 kV/m) [3]; thus; its installation does not represent a greater population risk, nor is it necessary to install an ADSS anti tracking cable because it presents potentials less than 10kV.

2) The contamination indexes where the installation of the ADSS fiber optic cable is projected, generally presents a level of light contamination levels I and II according to classification according to the International Electrotechnical Commission (IEC) TS 60815 -4: 2016, because they are:

- a) Zones without large industries do not produce high polluting smoke;
- b) Areas with low and/or moderate population density;
- c) Agricultural areas subject to frequent winds and rains;
- d) Areas distant from the sea > 10 km.

According to the reports obtained from the analysis of the electric field (potential gradient and electric potential) and environmental conditions of the fiber optic cable installation areas, it is argued that the evaluation of dry band arcing does not apply (dry band arcing), or corona discharges, as stipulated in the annexes of the IEEE 1222-2019 standard.

C. Evidence of Fiber Optic Installation

In Fig. 7, you can see the optical fiber installed on poles, where the recommendations set forth in the present investigation were followed, thus obtaining the installation of the optical fiber for the various geographical and climatological areas that owns the Lima-Peru region.



Fig. 7. Optical Fiber Installed in the Lima-Peru Region.

V. DISCUSSIONS

In [5], his research article Minimum technical specification of optical fiber for the construction of the National Backbone Network, he does not consider the climatic factors and geographical areas where the optical fiber will be installed, being a variable of interest for the present research work, considering that Peru has diverse and adverse geographic and climatic conditions, otherwise it occurs in the present research work, in which the geographic zones and climatological area are presented and described, where the optical fiber will be installed, due to the variant climate possessed by the various regions of Peru, especially the Lima region, which is part of the research analysis.

In [2], presents, in his research article Development of a software application to automate the design of optical links and its application in a link for the Minas - San Francisco, Ecuador power plant, the climatological factors are not considered and geographical areas in their development, this being a fundamental characteristic in the development of the current research, due to the fact that when transporting high information capacity, and even more so in the case of trunk fiber optic links, it is essential to consider these factors for the optimal functioning of the network, thus minimizing the cost of maintenance and loyalty of the service.

Taking into account the thesis [25], Design of a transport network over fiber optics to increase the broadband of the regions: Arequipa, Moquegua, Puno and Tacna, where he proposes the solution of enabling broadband, speeds higher than 256 kbit/s [26], on an existing overhead electrical network infrastructure, considering that the aforementioned regions are located in rugged geographical areas, areas with difficult to access and extreme climates, factors such as mechanical loads and electrical fields are not considered, which if they are considered as a fundamental part in the present research work, given that when transmitting high information capacity the installed or installed fiber optic has to overcome all geographical and environmental conditions.

As stated by [7], in his thesis, Design of a Broadband Telecommunications Network for the San Martín Region, he refers to the fact that the fiber optic cable will be installed on electricity network poles, however not describes the effects of

electric fields, nor the minimum safety distance, as recommended by the CNE [3]. The opposite is the case in the present research work, where if these essential factors are considered, such as the electric field produced by high, medium, and low voltage power lines, which minimizes the useful life of the optical fiber, in addition to safeguarding the Physical integrity of the engineers when carrying out the installation, the minimum safety distance between the electrical conductor and the optical fiber is considered.

Based on the citations of the previous research work of the region, no article was found that takes as a reference the CNE, for the development of their research proposals, where they describe and consider the area and geographical area, study of loads mechanical and electric fields as indispensable factors for the analysis and choice of the type of fiber optic cable to use for the installation, knowing that Lima region of Peru, has a great diversity of factors and climatic and geographical conditions.

VI. CONCLUSIONS

It is concluded after the research work carried out that the best option for choosing the type of optical fiber for installation in the Lima region, Peru, is the one that follows the ITU G.652.D standard, due to its special resistance characteristics to the mechanical loads and electric fields produced by the area, they do not affect the cable, thus minimizing maintenance costs and maximizing the durability, efficiency and quality of service to end users, generating greater profitability for companies, which unlike using cables of Optical fiber without conducting the aforementioned studies for the choice of cable type, does not guarantee durability or continuity of service and thus the loss of customer users.

It is concluded that based on the analysis and results obtained in the mechanical calculations of the fiber cable of the FiberHome company manufactured under the characteristics and recommendations of the ITU-T G.652.D, it complies with the maximum permissible stresses associated with the cable, the same that were given by the manufacturer, considering the geographic location, weather characteristics, wind pressure and ice, to be installed in zone B, areas 0, 1, 2 and 3, of the Lima region, Peru, on network lines electrical.

It is concluded that if all the companies that install fiber optic cables for high capacity links on existing medium and low voltage electrical network lines, would carry out the implementation under the recommendations indicated in the present research work, such as mechanical calculations of maximum allowable stress and electric fields, not only would increase the useful life of the optical fiber, but also the cost of maintenance would be reduced, thus avoiding breakdowns and failures in the operation of the network, which in many cases incur high costs for repair.

As a future research project, I recommend carrying out an analysis on the resistance to pollution of optical fibers according to the IEEE 1222-2019 standard for use in power lines, since the optical fiber when installed near environments of in where there are industrial areas, volcanic areas and areas with high salinity, particles such as sand and salt are deposited

on the surface of the cable and this could damage the optical fiber over time and affect operation.

REFERENCES

- [1] L. Hinojosa Gomez, "Selected Topics of Optical Fiber," Universidad Autónoma del Estado de Hidalgo, Pachuca-México, 2007.
- [2] W. Gómez Pauta, "Development of a software application to automate the design of optical links and their application in a link for the Minas - San Francisco power plant," Universidad del Azuay, Cuenca - Ecuador, 2017.
- [3] Dirección General de Electricidad, "National Code of Electricity-Utilization," Lima-Perú, 2006. Available: <http://intranet2.minem.gob.pe/web/cafae/Pdfs/CNE.PDF>.
- [4] L. S. Criollo Caizaguano, "Design of a convergent fiber optic network to interconnect the Campus of the University of the Americas," Quito / PUCE / 2015, Quito-Ecuador, 2017.
- [5] W. Azurza Neyra, "Minimum Technical Specification of Optical Fiber for the construction of the National Backbone Network," Lima-Perú, 2016. Available: <https://es.slideshare.net/wazurza/paper-especificacintcnica-mnima-de-fibra-ptica-para-la-construccin-de-la-red-dorsal-nacional>.
- [6] Furukawa Electric LatAm, "ADSS Optical Cables resistant to the tracking effect," Furukawa Electric, Dec. 22, 2017. <https://www.furukawalatam.com/es/conexion-furukawa-detalles/cables-opticos-adss-resistentes-al-efecto-tracking>.
- [7] Fiber Optic Association, "FOA Reference Guide To Fiber Optics," FOA, 2017. <https://www.foa.org/ESP/Cable.htm>.
- [8] L. Claudio Bedregal, "Design of a Broadband Telecommunications Network for the San Martín Region," Pontificia Universidad Católica del Perú, Lima-Perú, 2016.
- [9] International Telecommunication Union, "G.652 : Characteristics of a single-mode optical fibre and cable," ITU, 2017. <https://www.itu.int/rec/T-REC-G.652-201611-I/en>.
- [10] E. D. López Polo, "Design of a fiber optic network for the implementation of broadband service in Coishco (Ancash)," Universidad de Ciencias y Humanidades, Lima-Perú, 2016.
- [11] C. Barbut, "Next level for Fiber Optic Projects deployments," in Proceedings of the 11th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2019, Jun. 2019, pp. 1–3, doi: 10.1109/ECAI46879.2019.9042045.
- [12] C. Teneco, "FOA Reference Guide To Fiber Optics," FOA, 2017. https://foa.org/ESP/Fibra_optica.htm.
- [13] IEEE Standards Association, "1222-2019 - IEEE Standard for Testing and Performance for All-Dielectric Self-Supporting (ADSS) Fiber Optic Cable for Use on Electric Utility Power Lines," IEEE Xplore, Mar. 31, 2020. <https://ieeexplore.ieee.org/document/9052820>.
- [14] IEEE, "IEEE 1222-2019 - IEEE Standard for Testing and Performance for All-Dielectric Self-Supporting (ADSS) Fiber Optic Cable for Use on Electric Utility Power Lines," IEEE Standards Association, Nov. 07, 2019. <https://standards.ieee.org/content/ieee-standards/en/standard/1222-2019.html>.
- [15] Ministerio de Transportes y Comunicaciones, Technical Specifications for the Laying of Optical Fiber in Electric Power and Hydrocarbon Networks, Perú. Lima-Perú: Resolución Ministerial N° 368-2011-MTC/03, May 30, 2011, 2011, pp. 1–16.
- [16] FiberHome Technologies, "Optical Fiber & Cable," FiberHome, 2019. <http://www.fiberhome.com/>.
- [17] C. Barbut, "Fiber Optic Deployments in Romania between Metropolitan Fiber Optic Networks and Indoor Fiber Optic Infrastructure," in Proceedings of the 10th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2018, Apr. 2019, pp. 1–3, doi: 10.1109/ECAI2018.8679021.
- [18] C. Castillo, "Design of Transmission Lines: Don Bosco University | Electric power transmission | Distribution (commercial)," Soyapango-El Salvador, 2016. Available: <https://es.scribd.com/document/305312619/Clase-1-Diseno-de-Lineas-TX>.

- [19] ICNIRP International Commission on Non-Ionizing Radiation Protection, "Guidelines for limiting exposure to electromagnetic fields (100 kHz to 300 GHz)," Health Physics, May 01, 2020. <https://www.icnirp.org/en/activities/news/news-article/ef-guidelines-2020-published.html>.
- [20] P. Barcik, P. Munster, P. Dejdar, T. Horvath, and J. Vojtech, "Measurement of Polarization Transient Effects Caused by Mechanical Stress on Optical Fiber," in 2019 International Workshop on Fiber Optics in Access Networks, FOAN 2019, Sep. 2019, pp. 26–28, doi: 10.1109/FOAN.2019.8933658.
- [21] IEEE, "IEEE PC95.3 - IEEE Draft Recommended Practice for Measurements and Computations of Electric, Magnetic and Electromagnetic Fields with Respect to Human Exposure to Such Fields, 0 Hz-300 GHz," IEEE Standards Association, Feb. 06, 2016. https://standards.ieee.org/project/C95_3.html.
- [22] OFIL Systems, "ADSS Fiber Optic Cables in Overhead Transmission lines and the Corona phenomenon - OFIL - Daytime Corona Cameras," OFIL Systems, 2018. <https://ofilsystems.com/articles/adss-dry-band-arcing/>.
- [23] GL-Technology, "Opgw Cable Factory - China Opgw Cable Manufacturers & Suppliers," GL-Technology Manufacturer & Exporter, 2017. https://www.gl-fiber.com/products-opgw-cable/?gclid=Cj0KCQjwrsGCBhD1ARIsALILBYokSxfHoxtcvFyf-NqN-iut-ZpylxMAjjuI1MIUWQsAcYcNd7sh_IaAvKWEALw_wcB.
- [24] P. Taheri, B. Kordi, and A. M. Gole, "Electric field radiation from an overhead transmission line located above a lossy Ground," 2008, doi: 10.1109/UPEC.2008.4651634.
- [25] A. Nuñez Pacheco, "Design of a Transport Network on Optical Fiber to Increase the Broadband of the Regions: Arequipa, Moquegua, Puno and Tacna," Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú, 2018.
- [26] M. V. Alderete, "Broadband adoption in Latin American countries: does geographic proximity matter?," Probl. del Desarro. Rev. Latinoam. Econ., vol. 50, no. 198, pp. 31–56, May 2019, doi: 10.22201/ieec.20078951e.2019.198.67411.

On State-of-the-art of POS Tagger, “Sandhi” Splitter, “Alankaar” Finder and “Samaas” Finder for Indo-Aryan and Dravidian Languages

Hema Gaikwad¹, Jatinderkumar R. Saini²
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University)
Pune, India

Abstract—Computational Linguistic refers to the development of the computer systems that deal with human languages. In this paper, different Computational Linguistic Techniques such as Parts of Speech (POS) tagger, “Sandhi” Splitter, “Alankaar” Finder and “Samaas” Finder were considered. After a thorough literature review, it was found that fifteen techniques were used for POS tagging, nine techniques were used for “Sandhi” splitting, one work is done for “Alankaar” finder and absolutely no techniques are available for “Samaas” finder for the Indo-Aryan as well as Dravidian languages. Analysis shows that Rule Based Approach (RBA) and Hidden Markov Model (HMM) are frequently used for POS tagging, RBA is most frequently used for “Sandhi” Splitter, the general Human Intelligence (HI) is used for “Alankaar” Finder and no “Samaas” finder technique is available for any Indian language.

Keywords—“Alankaar”; “Samaas”; “Sandhi”; Parts of Speech tagger (POST)

I. INTRODUCTION

Natural Language Processing (NLP) has two main branches comprising of Natural Language Understanding (NLU) and Natural Language Generation (NLG). Computational Linguistic is a part of NLP and it requires a good understanding of both programming as well as knowledge of the language. Computational Linguistic techniques include Machine Translation, Speech Recognition systems, Text-to-Speech Synthesizers, Interactive Voice Response systems, Search Engines, POST, “Sandhi” Splitter, “Alankaar” Finder and “Samaas” Finder.

Many languages are spoken in different parts of India. The Indian languages can be divided mainly into Indo-Aryan and Dravidian languages. Punjabi, Hindi, Gujarati, Marathi, etc. are the examples of Indo-Aryan languages while Malayalam, Telugu, Kannada, etc. are the examples of the Dravidian languages. Hindi, recognized as the official language of India, is one of the most common languages in India [1]. It alone has 38 million native speakers and happens to be the fourth most spoken language of the world [2]. Hindi also has various dialects. For instance, Awadhi which is one of its dialects is spoken in 20 districts of India and 08 districts of Nepal [3]. The prominent texts like “Ramcharitmanas”, “Hanuman Chalisa” and “Padmavat” are written in Awadhi [4][5][6].

This paper presents a very thorough and exhaustive study of the various types of tools for the various Indian languages. The tools covered in this paper include POS tagger, “Sandhi” Splitter, “Alankaar” finder and “Samaas” finder. The best attempt has been made to present the research works done in the area till date.

The research work is segregated into various sections: Section II describes related work. Section III discusses the Analysis of NLP techniques for Indian Languages. Finally, the Section IV describes the Conclusion and Future work.

II. RELATED WORK

Basit et al. [7] talked about Awadhi POS tagger and its tag set. For developing tag set authors referred Bureau of Indian standards (BIS) and used Feature Based Approaches (FBA). Various features like word level, tag level, character level and Boolean level are used for POS tagging. Ekbal et al. [8] developed Bengali POS tagger using Maximum Entropy Approach (MEA). They worked on 72,341 words and uses 26 tags. MEA is based on feature selection and it can be lexicon feature, name entity recognition, suffix and prefix of word, context free feature, digit feature etc. By using the above features, the system got 88.2% accuracy. Proisl et al. [9] experimented parts of speech tagging on Magahi and Bhojpuri by using SoMeWeTa, Bi-Long Short Term Memory (LSTM)+Conditional Random Field (CRF) and Standard tagger approach. SoMeWeTa tagger depends on average structure perceptron. Bi-LSTM uses character word embedding and support transfer learning. Standard tagger based on Maximum Entropy Cyclic Dependency Network (MECDN). After experimenting, authors achieved 90.70% for Magahi and 94.08% for Bhojpuri. Ojha et al. [10] used CRF and Support Vector Machine (SVM) for tagging the Indo Aryan Languages Specifically Hindi, Odia and Bhojpuri. 90K tokens were used for training the system and 2K tokens were used for testing purpose. 88% to 93.7% accuracy was achieved with SVM and 82% to 86.7% accuracy was achieved with CRF.

Singh et al. [11] presented Bhojpuri POS tagger developed by SVM with 87.3% to 88.6% accuracy and errors can be minimized by increasing the corpus size. Pandey et al. [12] developed Chhattisgarhi POS tagger using RBA. 40,000 words (taken from story books) and 30 tags were used for testing purpose and achieved 78% accuracy. Sinha et al. [13]

presented Chhattisgarhi language rules so that this could further be used for developing parser and translators for the Chhattisgarhi language. Reddy et al. [14] developed cross language POS tagger using HMM i.e. Kannada POS tagger using Telugu resources. Bhirud et al. [15] talked about the significance of various Computational Linguistics (CL) tools such as Grammar checker, POST, “Sandhi” Splitter and “Samaas” Finder.

Verma et al. [16] talked about the Lexical analysis or tokenization process. The authors used different religious text such as Bible, Gita, Guru Granth Sahib, Rigveda and Quran to perform the lexical analysis process. Bhatt et al. [17] checked the accuracy of Gujrati POS tagger. For this, the author worked on two different data sets and two different methods. The data sets were Sports information dataset and Amusement dataset. By using HMM 70% and 56% accuracy were gained for sports information dataset and Amusement dataset respectively. By using RBA model, authors got 76% and 80% accuracy for sports information dataset and amusement dataset respectively. Sharma et al. [18] stated that multiple techniques were used to perform POS tagging on Hindi text. The techniques either based on Rules or based on Statistics or based on both. The statistical model could be SVM, HMM, CRF and MEA.

Narayan et al. [19] developed Hindi POS tagger using Artificial Neural Network (ANN) and achieved 91.03% accuracy. Narayan et al. [20] developed Hindi POS tagger using Quantum Neural Network (QNN) and achieved 99.13% accuracy. Mohnot et al. [21] proposed Hindi POS tagger developed using Hybrid Approach (HA) and it could be the combination of RBA, CRF, HMM and so on. 80,000 words and seven types of tags were used for experiment purpose. Joshi et al. [22] stated that three approaches were very common for POS tagging, they are RBA, Statistical Approach (SA) and HA. Garg et al. [23] used RBA for Hindi POS tagger. In this paper authors referred news, essay and short stories and collected 26,149 words and used 30 different tags and achieved 87.55% accuracy. Shrivastava et al. [24] developed Hindi POS tagger using Longest Suffix Matching Approach of HMM and got 93.12% accuracy. Dalal et al. [25] stated that Maximum Entropy Markov Model (MEMM) is used for POS tagging and chunking. This model is having various features such as corpus based feature, word based feature, dictionary based feature and context based features. The first three features are used for POS tagging and last feature is used for chunking purpose.

Antony et al. [26] developed Kannada POS tagger using SVM. Authors himself developed his own corpus, and words are taken from Kannada newspaper and books. Initially the corpus size was 1000 words then 25,000 words and finally 54,000 words and 30 tags. Accordingly, authors gained 48%, 66% and 86% accuracy respectively. Priyadarshi et al. [27] proposed Maithili POS using CRF. Author himself annotated Maithili text and created a corpus which consisted of 52,190 words. 85.88% accuracy was achieved when experiment was performed on wikipedia dumps and other Maithili web resources. Mundotiya et al. [28] developed Maithili POS tagger using CRF and achieved 0.77% precision & recall, 0.78% F1 score and 0.77% accuracy. Jha et al. [29] discussed about the “Sandhi” rules and Machine Learning models for analyzing the word, generating multiple words, concatenation with root word

to suffix or prefix. Singh et al. [30] developed morphology based Manipuri POS tagger. Authors used dictionaries for root word, prefix and suffix. System was tested on 3784 sentences that consist of 10,917 words. The result shows that 69% words were correctly tagged while 31% of them were incorrectly tagged (23% unknown words and 8% known words).

Patil et al. [31] developed Rule based Marathi POS tagger. The system is tested with small corpus size and achieved 78.82% accuracy. Authors stated that system’s accuracy can be increased by increasing the corpus size. Singh et al. [32] presented N-gram HMM for POS tagger. Authors considered tourism domain and collected 1,95,647 words for experiment purpose. Kaur et al. [33] talked about Punjabi POS tagger developed using HMM with tag set of 630 tags. Large tag set creates the data sparseness problem and it could be resolved by reducing the tag set. In this paper author suggested the new tag set proposed by Technical department of Indian languages (TDIL) and it consist of only 36 tags instead of 630 tags. The accuracy with 36 tags and 630 tags were 92-95% and 85-87% respectively. Mittal et al. [34] described N-gram HMM model for Punjabi POS tagger. Result showed that N-gram model is not suitable for unknown words because of spelling mistake or foreign language words.

Sharma et al. [35] stated that correctness of POS tagger depends on how accurately tagger tags the words of a sentence. The problem with the existing tagger is that it fails to tag the compound words and complex sentences. Authors were interested to increase the efficiency of existing Punjabi POS tagger by implementing the Viterbi algorithm of Bi-gram HMM. Suresh et al. [36] developed Telugu POS tagger using HMM with 620 tags but TDIL proposed only 34 tags for Indian languages. After experimenting 92-95% and 85-87% accuracy achieved with 34 tags and 620 tags respectively. Jagadeesh et al. [37] used unsupervised learning algorithm and Deep Learning (DL) methods for developing Telugu POS. The Table I indicate approaches for POS tagger for Indian Languages.

Al Shamsi et al. [38] used HMM to develop Arabic POS tagger and got 97% accuracy. Demilie et al. [39] developed POS tagger for Awngi language using HMM. Authors used 23 tags and 188,760 words for training and testing purpose. 93.64% and 94.77% accuracy is achieved with Uni-gram and Bi-gram HMM respectively. Purnamasari et al. [40] talked about Indonesian rule based POS tagger and authors used KBBI (Indonesian large dictionary) and morphological rules for tagging purpose.

Wicaksono et al. [41] developed POS tagger for Indonesian language using HMM. Affix tree, succeeding POS tag and additional lexicon methods were used to improve the accuracy. The result stated that affix tree and additional lexicon methods are best to improve the accuracy of POS tagger than succeeding POS tag. Dibitso et al. [42] developed Setswana African Language POS using SVM. Authors reviewed POS taggers for different African languages and identified challenges and techniques. Table II shows approaches for POS tagger for International Languages from 2006 to 2019 duration.

TABLE I. APPROACHES FOR POS TAGGER FOR INDIAN LANGUAGES

S.No	Author(s)	Language	Year	Approach
1	Basit et al. [7]	Awadhi	2008	FBA
2	Ekbal et al. [8]	Bengali	2008	MEA
3	Proisl et al. [9]	Bhojpuri & Magadhi	2019	SoMeWeTa, MECDN, Bi-LSTM+CRF
4	Ojha et al.[10]	Bhojpuri	2015	SVM, CRF
5	Singh et al.[11]	Bhojpuri	2015	SVM
6	Pandey et al. [12]	Chhattisgarhi	2018	RBA
7	Sinha et al. [13]	Chhattisgarhi	2018	RBA
8	Reddy et al. [14]	Cross Language	2011	HMM
9	Bhirud et al. [15]	Generic	2017	CL
10	Verma et al. [16]	Generic	2017	ML
11	Bhatt et al. [17]	Gujrati	2019	HMM, RBA
12	Sharma et al. [18]	Hindi	2020	RBA, SA, HA
13	Narayan et al. [19]	Hindi	2014	ANN
14	Narayan et al. [20]	Hindi	2014	QNN
15	Mohnot et al. [21]	Hindi	2014	HA
16	Joshi et al. [22]	Hindi	2013	HMM
17	Garg et al. [23]	Hindi	2012	RBA
18	Shrivastava et al. [24]	Hindi	2008	HMM
19	Dalal et al. [25]	Hindi	2006	MEMM
20	Antony et al. [26]	Kannada	2010	SVM
21	Priyadarshi et al. [27]	Maithili	2020	CRF
22	Mundotiya et al. [28]	Maithili	2020	CRF
23	Jha et al. [29]	Maithili	2018	RBA
24	Singh et al. [30]	Manipuri	2008	MBA
25	Patil et al. [31]	Marathi	2014	RBA
26	Singh et al. [32]	Marathi	2013	N-gram HMM
27	Kaur et al. [33]	Punjabi	2015	HMM
28	Mittal et al. [34]	Punjabi	2014	HA
29	Sharma et al. [35]	Punjabi	2011	Bi-gram HMM
30	Suresh et al. [36]	Telugu	2019	HMM
31	Jagadeesh et al. [37]	Telugu	2016	DL

TABLE II. APPROACHES FOR POS TAGGER FOR INTERNATIONAL LANGUAGES

S.No	Author(s)	Language	Year	Approach
1	Al Shamsi et al. [38]	Arabic	2006	HMM
2	Demilie et al. [39]	Awangi	2019	HMM
3	Purnamasari et al. [40]	Indonesian	2018	RBA
4	Wicaksono et al. [41]	Indonesian	2010	HMM
5	Dibitso et al. [42]	Setswana African	2019	SVM

Kovida et al. [43] discussed General Approaches (GA) used for language independent “Sandhi” Splitter and the system has been tested on two languages Telugu and Malayalam. Devadath et al. [44] conducted “Sandhi” splitting experiment on Dravidian languages. Authors evaluated the performance of “Sandhi” splitting tool and analyzed error propagation rate. Joshi et al. [45] presented “Sandhi” viched (“Sandhi” Splitter) using different Hindi rules. They experimented their system on 847 Hindi compound words and got 75% accuracy. Gupta et al. [46] developed a Rule based “Sandhi” Viched system for Hindi Language. The authors tested the system on more than 200 words and got 60% to 80% accuracy. Deshmukh et al. [47] compared four “Sandhi” analyzer and “Sandhi” Splitter systems developed in Sanskrit, Marathi, Hindi and Malayalam and authors found that RBA was used for all four languages.

Murthy et al. [48] developed first “Sandhi” Splitter in Kannada using “Sandhi” Place Determination (SPD) and Prefix Suffix method (PSM). The experiment was performed on 7000 words in Kannada language and achieved 80% accuracy. Shashirekha et al. [49] presented RBA based agama “Sandhi” Splitter namely Yakaragama and Vakaragama. The experiment was tested on the words taken from Kannada newspaper and online resources. The developed system achieved 98.85% accuracy.

Shree et al. [50] proposed Kannada “Sandhi” Splitter using CRF method. Sebastian et al. [51] discussed the results and issues of Malayalam word Splitter developed using Machine Learning (ML) approaches. Premjith et al. [52] used DL methods such as RNN, LSTM and Gated Recurrent Units (GRU) for constructing and splitting the words and obtained 98.08%, 97.88% and 98.16% accuracy respectively. Nisha et al. [53] developed the Malayalam “Sandhi” Splitter using Memory Based Language Processing (MBLP) algorithm. This algorithm was based on suffix separation. Authors discussed three methods for suffix separation such as Root driven method, Affix stripping method and the Suffix stripping method. Devadath et al. [54] developed the Malayalam “Sandhi” Splitter using the HA and got 91.1% accuracy and authors stated that HA was better than RBA and SA, because it is faster and more accurate.

Das et al. [55] developed Malayalam “Sandhi” Splitter using HA and Malayalam characters were represented by unicode. Nair et al. [56] developed Malayalam “Sandhi” Splitter using RBA to split the compound words. The system was tested on 4000 compound words and got 90% accuracy. Authors stated that work can be extended to other Dravidian languages because they have structural similarity. Joshi et al. [57] developed Marathi “Sandhi” Splitter using RBA. The experiment was tested on 150 words and got 70-80% accuracy. Patil et al. [58] proposed “Sandhi” viched system for Sanskrit language using RBA.

Bhardwaj et al. [59] developed Sanskrit benchmark called “Sandhi”kosh. “Sandhi”kosh includes Rule based corpus, Literature corpus, Bhagavad Gita corpus, UoH corpus and Astaadhyayi. In this paper authors presented three most popular “Sandhi” splitting tools such as JNU tool, UoH tool and INRIA tool. All these tools refer “Sandhi”kosh for

referring any rules. All these are openly available and can be used by anyone for validating their tools.

Hellwig et al. [60] introduced Convolution Neural Network (CNN) and RNN for splitting the Sanskrit compound words and this model is also suitable for German compound words. Hellwig et al. [61] developed “Sandhi” resolution and “Sandhi” splitting system using RNN. Natarajan et al. [62] used Bayesian Word Segmentation Method (BWSM) for Sanskrit “Sandhi” Splitter. Rao et al. [63] focused Consonant and Phrase based “Sandhi” splitting for Telugu language. Vempaty et al. [64] developed a “Sandhi” Splitter for Telugu language by using Finite State Automata (FSA). The corpus size is 158K words and authors got 80.30% accuracy on 500 words. Table III depicts approaches for “Sandhi” Splitter for Indian Languages.

Adhikari et al. [65] discussed the rules for improving the existing Nepali morphological analyzers. Paul et al. [66] discussed about the Nepali stemmer developed using an affix stripping technique and rule based technique. The system was tested on 1800 words of different domain. These domains include news on Economics, Health & Political in Nepali language, which are based on Devanagari Script. The overall accuracy of the designed system was 90.48%. Basapur et al. [67] stated that developing a “Sandhi” Splitter or “Sandhi” joiner for Pali language is bit difficult because the complex nature of grammar rules. The Table IV represents approaches for “Sandhi” Splitter for International Languages from 2014 to 2020.

Hemlata et al. [68] stated that translation is the process of changing the words from one language to the other language without altering the meaning. Translation is a difficult task because it involves large no. of Ras and Alankaar. These help to enhance the beauty of the literature. Ramcharitmanas is an Awadhi epic which has a tremendous usage of Alankaar. It can be translated through machine, but doing so will deplore the beauty of the epic. Authors did this work better with the help of Human Intelligence (HI).

Das et al. [69] stated parse structure and simple sentence generation algorithm are used to generate simple sentences from the complex or compound sentences. Sharma [70] stated two things. Firstly, sentence simplification methods are used to simplify compound sentences. Secondly the RBA, HMM POS tagger and lexicon based morph are used to identify syntactic errors. On testing, the system got 93.30% precision, 97.32% recall rate and 95.25% F measures. Garain et al. [71] stated that sentences can be simplified by preparing parse tree and their efficiency could be decided on the basis of parse tree’s efficiency. Poornima et al. [72] defined the RBA for sentence simplification. It is a two-step process. In first step, split the sentence by seeing the delimiter and in second step again split the sentence by seeing the connectives. Zhu et al. [73] stated that sentence simplification process consists of source and target. Complex sentence and simple sentence could be source and target. Tree based simplification model is used for splitting, dropping, reordering and substitution.

As discussed above, although some papers on sentence simplification were found, no papers were found on “Samaas” Finder for any Indian language.

TABLE III. APPROACHES FOR “SANDHI” SPLITTER FOR INDIAN LANGUAGES

S.No	Author(s)	Language	Year	Approach
1	Kovida et al [43]	Agglutinative Language	2011	GA
2	Devadath et al [44]	Dravidian	2016	GA
3	Joshi et al [45]	Hindi	2016	RBA
4	Gupta et al [46]	Hindi	2009	RBA
5	Deshmukh et al [47]	Indian Language	2014	GA
6	Murthy et al[48]	Kannada	2017	SPD, PSM
7	Shashirekha et al[49]	Kannada	2016	RBA
8	Shree et al[50]	Kannada	2016	CRF
9	Sebastian et al [51]	Malayalam	2020	ML
10	Premjith et al [52]	Malayalam	2018	DL
11	Nisha et al [53]	Malayalam	2016	MBLP
12	Devadath et al [54]	Malayalam	2014	RBA, SA
13	Das et al [55]	Malayalam	2012	RBA,ML
14	Nair et al [56]	Malayalam	2011	RBA
15	Joshi et al [57]	Marathi	2012	RBA
16	Patil et al [58]	Sanskrit	2018	RBA
17	Bhardwaj et al [59]	Sanskrit	2018	RBA
18	Hellwig et al [60]	Sanskrit	2018	ANN, QNN
19	Hellwig et al [61]	Sanskrit	2015	RNN
20	Natarajan et al [62]	Sanskrit	2011	BWSM
21	Rao et al [63]	Telugu	2014	RBA
22	Vempaty et al [64]	Telugu	2011	FSA

TABLE IV. APPROACHES FOR “SANDHI” SPLITTER FOR INTERNATIONAL LANGUAGES

S.No	Year	Author(s)	Language	Approach
1	2020	Adhikari et al. [65]	Nepali	RBA
2	2014	Paul et al. [66]	Nepali	RBA
3	2019	Basapur et al. [67]	Pali	GA

III. ANALYSIS OF NLP TECHNIQUES FOR INDIAN LANGUAGES

After analyzing the contents of Table I and Table III, we find that fifteen techniques are used for POS tagging and nine techniques are used for “Sandhi” splitting for many Indian Languages. Very less work is done for “Alankaar” Finder and no work is done for “Samaas” finder. Table V indicates POS tagger approaches abbreviation, Table VI represents “Sandhi” Splitter approaches abbreviation and Table VII indicates “Alankaar” Finder approach abbreviation.

Various graphs have been prepared by considering the different parameters. Fig. 1 shows Language-wise available POS tagger. Fig. 2 is for No. of approaches used by POS tagger and Fig. 3 is year-wise POS tagger.

TABLE V. POS TAGGER APPROACHES AND ITS ABBREVIATION

S.No	Approach name	Abbreviation
1	Rule Based Approach	RBA
2	Stochastic Approach	SA
3	Hybrid Approach	HA
4	Artificial Neural Network	ANN
5	Quantum Neural Network	QNN
6	Hidden Markov Model	HMM
7	Maximum Entropy Markov Model	MEMM
8	N Gram Markov Model	NGMM
9	Feature Based Approach	FBA
10	Conditional Random Field	CRF
11	Support Vector Machine	SVM
12	Morphology Based Approach	MBA
13	Author has not provided the details	SoMeWeTa
14	Bi-Long Short Term Memory	Bi-LSTM
15	Maximum Entropy Cyclic Dependency Network	MECDN

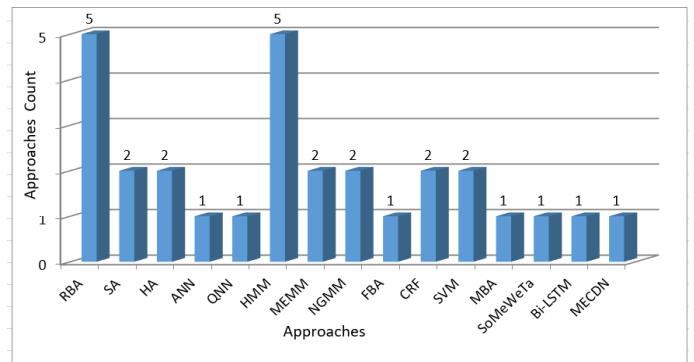


Fig. 2. No. of Approaches used by POS Taggers.

TABLE VI. "SANDHI" SPLITTER APPROACHES AND ITS ABBREVIATION

S.No	Approach name	Abbreviation
1	Rule Based Approach	RBA
2	Deep Learning	DL
3	Machine Learning	ML
4	Conditional Random Field	CRF
5	Memory Based Language Processing	MBLP
6	Bayesian Word Segmentation Method	BWSM
7	Finite State Automata	FSA
8	"Sandhi" Place Determination	SPD
9	Prefix and Suffix Method	PSM

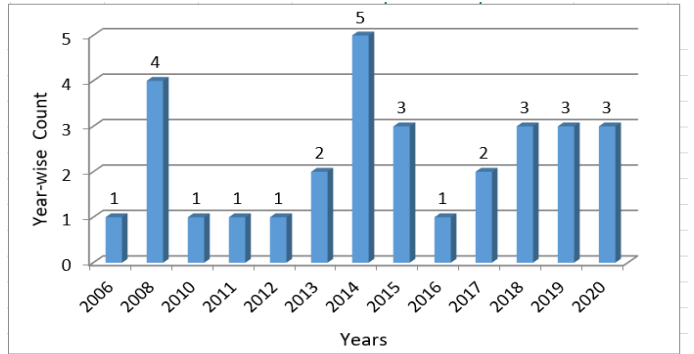


Fig. 3. Year-Wise POS Tagger.

Different graphs have been made for "Sandhi" Splitter. Fig. 4 represent language-wise available "Sandhi" Splitter. Fig. 5 shows the No. of approaches used by "Sandhi" Splitter and Fig. 6 is year-wise "Sandhi" Splitter.

TABLE VII. "ANANKAAR" FINDER APPROACH AND ITS ABBREVIATION

S.No	Approach name	Abbreviation
1	Human Intelligence	HI

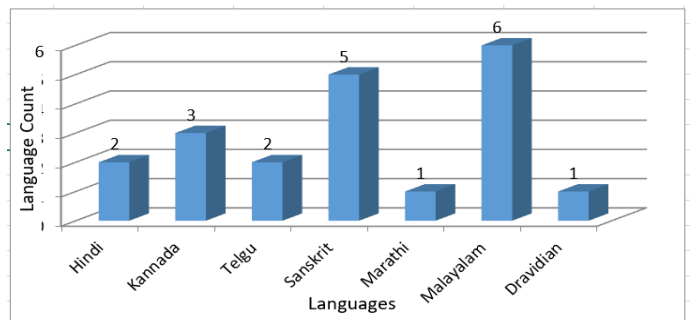


Fig. 4. Language-Wise Available "Sandhi" Splitter.

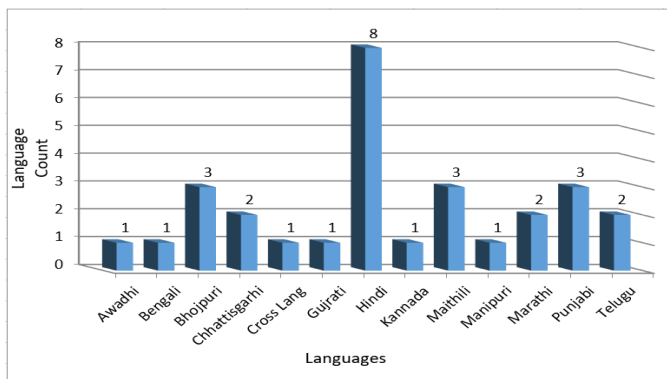


Fig. 1. Language-Wise Available POS Tagger.

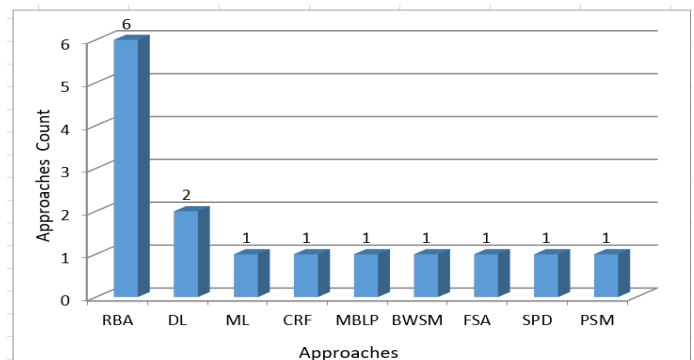


Fig. 5. No. of Approaches used by "Sandhi" Splitter.

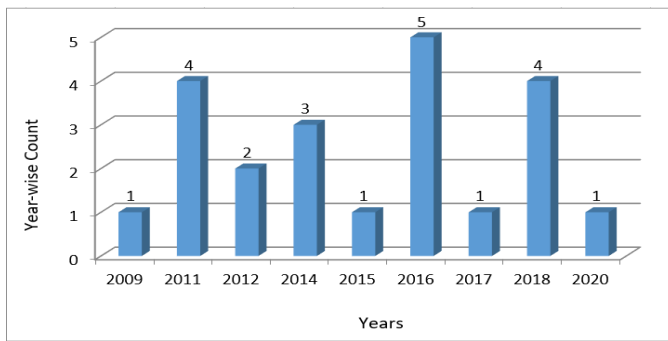


Fig. 6. Year-Wise "Sandhi" Splitter.

Fig. 7 shows the various approaches used by different Computational Linguistic tools.

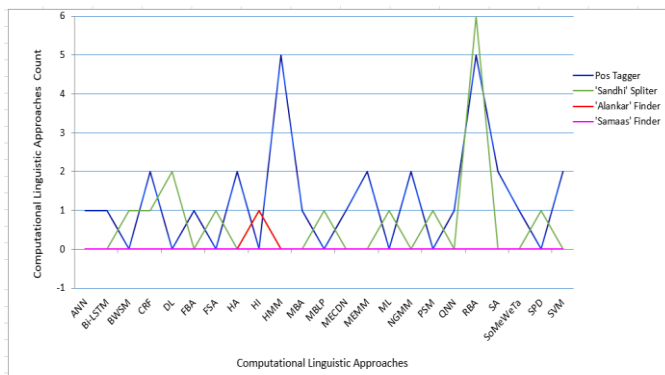


Fig. 7. Different Approaches used by POS Tagger, "Sandhi" Splitter, "Alankaar" Finder and "Samaas" Finder.

After reviewing all research papers, it is observed that most of the Computational Linguistics work is done in Maharashtra, Punjab, Telangana, Tamil Nadu and Uttar Pradesh. Table VIII depicts the state wise statistics.

Fig. 8 shows the Political map of India [74] and the state wise linguistic work are represented on the map.

TABLE VIII. COMPUTATIONAL LINGUISTICS STATISTICS STATE WISE

State	Count	State	Count
Andhra Pradesh	2	Madhya Pradesh	1
Assam	1	Maharashtra	8
Bihar	1	Meghalaya	1
Chhattisgarh	3	Punjab	7
Delhi	3	Rajasthan	2
Gujrat	1	Tamil Nadu	4
Haryana	1	Telangana	5
Jharkhand	1	Uttar Pradesh	4
Karnataka	3	West Bengal	3
Kerala	4		

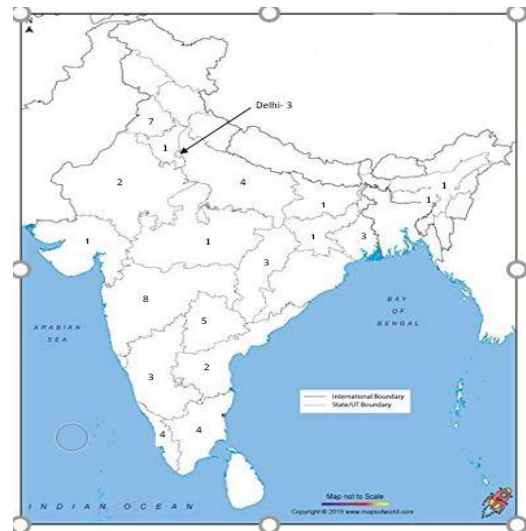


Fig. 8. State Wise Computational Linguistic Work.

IV. CONCLUSION AND FUTURE WORK

Linguistic techniques are helpful for understanding the natural languages. Four Computational Linguistic tools namely POS tagger, "Sandhi" Splitter, "Alankaar" Finder and "Samaas" Finder for Indo-Aryan and Dravidian languages have been considered. It is observed that POS tagger and "Sandhi" Splitter are available while "Alankaar" Finder and "Samaas" Finder are not. Most of the POS taggers are available only for Hindi language while "Sandhi" splitters are available mostly for Malayalam language. Fifteen techniques such as RBA, SA, HA, ANN, QNN, HMM, MEMM, N-gram HMM, FBA, CRF, SVM, MBA, Bi-LSTM and MECDN are suitable for POS tagging. It is observed that most of the Indian language POS taggers are built by using RBA and HMM.

Nine techniques namely RBA, DL, ML, CRF, MBLP, BWSM, FSA, SPD and PSM are appropriate for "Sandhi" Splitter. RBA is commonly used by researchers for developing "Sandhi" Splitter. The study shows that HI could be used for "Alankaar" Finder. But technique for "Samaas" Finder are unavailable yet.

As a future work, the authors would like to extend this work and use ML techniques for linguistic tools i.e. POS tagger, "Sandhi" Splitter, "Alankaar" Finder and "Samaas" Finder for Indo-Aryan and Dravidian languages.

REFERENCES

- [1] The Constitution Of India, (2019). Government Of India Ministry Of Law and Justice Legislative Department.
- [2] Information from Omniglot (2008), The online encyclopedia of writing systems and languages.
- [3] Internet Archive, (2007), Linguistic Survey of India Vol. 6.
- [4] Ramcharitmans, (2015), Tulsidas Ramcharitmanas.
- [5] The Divine India, (2020), Hanuman Chalisa.
- [6] KavitaKosh, (2006), Ramcharitmans/Tulsidas.
- [7] Basit, A., & Kumar, R. (2019) Towards a Part-of-Speech Tagger for Awadhi: Corpus and Experiments.

- [8] Ekbal, A., Haque, R., & Bandyopadhyay, S. (2008). Maximum entropy based Bengali part of speech tagging. *Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*, 33, 67-78.
- [9] Proisl, T., Uhrig, P., Blombach, A., Dykes, N., Heinrich, P., Kabashi, B., & Mammarella, S. (2019). The Illiterati: Part-of-Speech Tagging for Magahi and Bhojpuri without even knowing the alphabet. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers* (pp. 73-79).
- [10] Ojha, A. K., Behera, P., Singh, S., & Jha, G. N. (2015). Training & evaluation of POS taggers in Indo-Aryan languages: a case of Hindi, Odia and Bhojpuri. In the proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (pp. 524-529).
- [11] Singh, S., & Jha, G. N. (2015, August). Statistical tagger for Bhojpuri (employing support vector machine). In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1524-1529). IEEE.
- [12] Pandey, V., Padmavati, M. V., & Kumar, R. Rule Based Parts of Speech Tagger for Chhattisgarhi Language.
- [13] Sinha, S.K. Sahu, & S ther (2018). Parts of speech tagging for Chhattisgarhi language. *International journal of creative research thoughts* (Volume 6, Issue 1 February 2018, ISSN: 2320-2882).
- [14] Reddy, S., & Sharoff, S. (2011, November). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proceedings of the Fifth International Workshop On Cross Lingual Information Access* (pp. 11-19).
- [15] Bhirud, N. S., Bhavsar, R., & Pawar, B. (2017). Grammar checkers for natural languages: a review. *International Journal on Natural Language Computing (IJNLC)*, 6(4), 51-62.
- [16] Verma, M. (2017). Lexical analysis of religious texts using text mining and machine learning tools. *International Journal of Computer Applications*, 168(8), 39-45.
- [17] Bhatt, P. M., & Ganatra, A. Analyzing& enhancing accuracy of part of speech tagger with the usage of mixed approaches for Gujarati. *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277, 3878.
- [18] Sharma, A., & Yadav, V. Approaches to Part of speech Tagging in Hindi Language: A Review.
- [19] Narayan, R., Chakraverty, S., & Singh, V. P. (2014). Neural network based parts of speech tagger for Hindi. *IFAC Proceedings Volumes*, 47(1), 519-524.
- [20] Narayan, R., Singh, V. P., & Chakraverty, S. (2014). Quantum neural network based parts of speech tagger for Hindi. *International Journal of Advancements in Technology*, 5(2), 137-152.
- [21] Mohnot, K., Bansal, N., Singh, S. P., & Kumar, A. (2014). Hybrid approach for Part of Speech Tagger for Hindi language. *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, 4(1), 25-30.
- [22] Joshi, N., Darbari, H., & Mathur, I. (2013). HMM based POS tagger for Hindi. In *Proceeding of 2013 International Conference on Artificial Intelligence, Soft Computing (AISC-2013)* (pp. 341-349).
- [23] Garg, N., Goyal, V., & Preet, S. (2012, December). Rule based Hindi part of speech tagger. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 163-174).
- [24] Shrivastava, M., & Bhattacharyya, P. (2008, December). Hindi POS tagger using naive stemming: harnessing morphological information without extensive linguistic knowledge. In *International Conference on NLP (ICON08)*, Pune, India.
- [25] Dalal, A., Nagaraj, K., Sawant, U., & Shelke, S. (2006). Hindi part-of-speech tagging and chunking: A maximum entropy approach. *Proceeding of the NLP/ML Machine Learning Competition*.
- [26] Antony, P. J., & Soman, K. P. (2010, July). Kernel based part of speech tagger for Kannada. In *2010 International Conference on Machine Learning and Cybernetics (Vol. 4, pp. 2139-2144)*. IEEE.
- [27] Priyadarshi, A., & Saha, S. K. (2020). Towards the first Maithili part of speech tagger: Resource creation and system development. *Computer Speech & Language*, 62, 101054.
- [28] Mundotiya, R. K., Singh, M. K., Kapur, R., Mishra, S., & Singh, A. K. (2020). Basic Linguistic Resources and Baselines for Bhojpuri, Magahi and Maithili for Natural Language Processing. *arXiv preprint arXiv:2004.13945*.
- [29] Jha, S. K., Singh, P. P., & Kaul, V. K. (2018). VEA Model in Word Formation Process of Maithili MT.
- [30] Singh, T. D., & Bandyopadhyay, S. (2008). Morphology driven manipurios tagger. In *Proceedings of the IJCNLP-08 Workshop on NLP for less privileged languages*.
- [31] Patil, H. B., Patil, A. S., & Pawar, B. V. (2014). Part-of-Speech Tagger for Marathi Language using Limited Training Corpora. *International Journal of Computer Applications*, 975, 8887.
- [32] Singh, J., Joshi, N., & Mathur, I. (2013, August). Development of Marathi part of speech tagger using statistical approach. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1554-1559). IEEE.
- [33] Kaur, M., Aggerwal, M., & Sharma, S. K. (2014). Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set. *International Journal of Computer Applications & Information Technology*, 7(2), 142.
- [34] Mittal, S., Sethi, N. S., & Sharma, S. K. (2014). Part of speech tagging of Punjabi language using N gram model. *International Journal of Computer Applications*, 100(19).
- [35] Sharma, S. K., & Lehal, G. S. (2011, June). Using hidden markov model to improve the accuracy of Punjabi pos tagger. In *2011 IEEE International Conference on Computer Science and Automation Engineering (Vol. 2, pp. 697-701)*. IEEE.
- [36] Suresh, V. Reduced Tagset To Improve Accuracy of HMM Based Parts of Speech Tagger in Telugu Language.
- [37] Jagadeesh, M., Kumar, M. A., & Soman, K. P. (2016). Deep belief network based part-of-speech tagger for Telugu language. In *Proceedings of the Second International Conference on Computer and Communication Technologies* (pp. 75-84). Springer, New Delhi.
- [38] Al Shamsi, F., & Guessoum, A. (2006, April). A hidden Markov model-based POS tagger for Arabic. In *Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France* (pp. 31-42).
- [39] Demilie, W. B. (2019, September) Parts of Speech Tagger for Awngi Language. *International journal of Engineering Science and Computing (Vol. 9, Issue No 9)*.
- [40] Purnamasari, K. K., & Suwardi, I. S. (2018, September). Rule based Part of Speech Tagger for Indonesian Language. In *IOP Conference Series: Materials Science and Engineering (Vol. 407, No. 012151, pp. 1-4)*.
- [41] Wicaksono, A. F., & Purwarianti, A. (2010, August). HMM Based part-of-speech tagger for bahasa Indonesia. In *Fourth International MALINDO Workshop, Jakarta*.
- [42] Dibitso, M. A., Owolawi, P. A., & Ojo, S. O. (2019, November). Part of Speech Tagging for Setswana African Language. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)* (pp. 1-6). IEEE.
- [43] Kovida, K. P. N., Sneha, N., & Mamidi, R. (2011). Statistical Sandhi Splitter For Agglutinative Languages.
- [44] Devadath, V. V., & Sharma, D. M. (2016, August). Significance of an accurate sandhi-Splitter in shallow parsing of dravidian languages. In *Proceedings of the ACL 2016 Student Research Workshop* (pp. 37-42).
- [45] Joshi, B. K., & Kushwah, K. K. (2016). Sandhi: the rule based word formation in Hindi. *International Journal of Computer Science and Information Security*, 14(12), 781.
- [46] Gupta, P., & Goyal, V. (2009). Implementation of rule based algorithm for Sandhi-Viched of compound Hindi words. *arXiv preprint arXiv:0909.2379*.
- [47] Deshmukh, R., Bhojane, V., & PIIT, N. P. (2014). Sandhi Splitting Techniques For Different Indian Languages. *International Journal of Engineering Technology, Management and Applied Sciences (ijetmas)*, 2(7).

- [48] Murthy, S. R., Akshatha, A. N., Upadhyaya, C. G., & Kumar, P. R. (2017, September). Kannada spell checker with sandhi Splitter. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 950-956). IEEE.
- [49] Shashirekha, H. L., & Vanishree, K. S. (2016, September). Rule based Kannada Agama Sandhi Splitter. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 549-553). IEEE.
- [50] Shree, M. R., Lakshmi, S., & Shambhavi, B. R. (2016, October). A novel approach to Sandhi splitting at character level for Kannada language. In 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS) (pp. 17-20). IEEE.
- [51] Sebastian, M. P., & Kumar, G. S. (2020). Machine learning approach to suffix separation on a sandhi rule annotated malayalam data set. *South Asia Research*, 40(2), 231-249.
- [52] Premjith, B., Soman, K. P., & Kumar, M. A. (2018). A deep learning approach for Malayalam morphological analysis at character level. *Procedia computer science*, 132, 47-54.
- [53] Nisha, M., & Raj, P. R. (2016). Sandhi Splitter for malayalam using mb1p approach. *Procedia Technology*, 24, 1522-1527.
- [54] Devadath, V. V., Kurisinkel, L. J., Sharma, D. M., & Varma, V. (2014, December). A sandhi Splitter for malayalam. In Proceedings of the 11th International Conference on Natural Language Processing (pp. 156-161).
- [55] Das, D., Radhika, K. T., Rajeev, R. R., & PC, R. R. (2012). Hybrid sandhi-Splitter for malayalam using unicode. In in proceedings of National Seminar on Relevance of Malayalam in Information Technology.
- [56] Nair, L. R., & Peter, S. D. (2011, September). Development of a rule based learning system for splitting compound words in Malayalam language. In 2011 IEEE Recent Advances in Intelligent Computational Systems (pp. 751-755). IEEE.
- [57] Joshi Shripad, S. (2012). Sandhi splitting of Marathi compound words. *Int. J. on Adv. Computer Theory and Engg*, 2(2).
- [58] Patil, B., & Patil, M. (2018). Implementation of Sandhi Viccheda for Sanskrit Words/Sentences/Paragraphs.
- [59] Bhardwaj, S., Gantayat, N., Chaturvedi, N., Garg, R., & Agarwal, S. (2018, May). Sandhikosh: A benchmark corpus for evaluating sanskrit sandhi tools. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
- [60] Hellwig, O., & Nehrlich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2754-2763).
- [61] Hellwig, O. (2015, November). Using Recurrent Neural Networks for joint compound splitting and Sandhi resolution in Sanskrit. In 4th Biennial Workshop on Less-Resourced Languages.
- [62] Natarajan, A., & Charniak, E. (2011, November). S3-Statistical Sandhi Splitting-Statistical Sandhi Splitting. In Proceedings of 5th International Joint Conference on Natural Language Processing (pp. 301-308).
- [63] Rao, T. K., & Prasad, T. V. (2014). Telugu Bigram Splitting using Consonant-based and Phrase-based Splitting. *Editorial Preface*, 5(5), 122.
- [64] Vempaty, P. C., & Nagalla, S. C. P. (2011). Automatic sandhi splitting method for Telugu, an Indian language. *Procedia-Social and Behavioral Sciences*, 27, 218-225.
- [65] Adhikari, M., & Neupane, A. (2020). A vowel based word Splitter to improve performance of existing Nepali morphological analyzers on words borrowed from Sanskrit. *Kathmandu University Journal of Science, Engineering and Technology*, 14(1).
- [66] Paul, A., Dey, A., & Purkayastha, B. S. (2014). An Affix Removal Stemmer for Natural Language Text in Nepali. *International Journal of Computer Applications*, 91(6).
- [67] Basapur, S., Shivani, V., & Nair, S. (2019). Pāli Sandhi—A computational approach. In Proceedings of the 6th International Sanskrit Computational Linguistics Symposium (pp. 181-192).
- [68] Hemlata, M. A., & Kalan, B. K. Distortion or Translation (2019). *Studying Figures of Speech in Ramcharitmanasa*.
- [69] Das, B., Majumder, M., & Phadikar, S. (2018). A novel system for generating simple sentences from complex and compound sentences. *International Journal of Modern Education and Computer Science*, 12(1), 57.
- [70] Sharma, S. K. (2019). Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language. *EAI Endorsed Transactions on Scalable Information Systems*, 6(20), e4.
- [71] Garain, A., Basu, A., Dawn, R., & Naskar, S. K. (2019, November). Sentence simplification using syntactic parse trees. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON) (pp. 672-676). IEEE.
- [72] Poornima, C., Dhanalakshmi, V., Anand, K. M., & Soman, K. P. (2011). Rule based sentence simplification for english to tamil machine translation system. *International Journal of Computer Applications*, 25(8), 38-42.
- [73] Zhu, Z., Bernhard, D., & Gurevych, I. (2010, August). A monolingual tree-based translation model for sentence simplification. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010) (pp. 1353-1361).
- [74] Maps of Inida, (2020), India Outline Map with States and Union Territories.

The Development of Students' Spatial Orientation Through the Use of 3D Graphics

Benjamín Maraza-Quispe

Facultad de Ciencias de la Educación

Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú

Abstract—The purpose of this research is to determine to what extent the use of 3D graphics in the educational process improves the spatial orientation skills of secondary school students. The research follows a qualitative approach of experimental type, the population is constituted by 300 students of Secondary Education of which through a simple random sampling 25 students were chosen. Four sessions of 50 minutes each have been developed, in which three-dimensional models were used, in order to determine if spatial skills are developed. A psychometric pre-test and post-test of spatial reasoning was taken in order to determine how much the spatial skills of the selected sample members are developed based on the measurement of five criteria: Construction of three-dimensional objects (intermediate level), Construction of three-dimensional objects (advanced level), Rotation of objects from references (intermediate level), Rotation of objects from references (advanced level) and deconstruction of three-dimensional objects. For the data analysis, the data from the scores obtained by the students in both the pre-test and the post-test are processed. The results allow us to visualize that the use of 3D graphics in the teaching-learning processes allows us to improve spatial orientation skills to a great extent. The result is evidenced in the increase of the total scores obtained in the post-test in comparison with the results of the pre-test. Likewise, an increase from 47.9% to 75.1% of items answered correctly was observed on average, which was corroborated with the Student's t-test that gave a P value of less than 0.05, demonstrating the reliability of the research developed and therefore significantly improving spatial orientation skills in students through the use of 3D graphics technology.

Keywords—Orientation; reasoning; spatial; technology; 3D graphics; education; processes; educational

I. INTRODUCTION

Technology is developing by leaps and bounds, expanding to several areas, such as education, where more and better technological tools are required to facilitate the learning process of students. The research revolves around the question, to what extent does the use of 3D graphics technology allow the development of spatial orientation in secondary school students?

The importance of the research lies in the impact of the use of 3D graphics in the field of education, allowing students to develop their spatial orientation, which will be useful to have a better performance in a given space and at the same time promote the use of 3D technology in educational institutions, since, the main problem of many of them and education in general is that the use of new technologies have not been

widely developed, on the contrary, monotonous and routine teaching models are still used, leading to student fatigue, as well as low academic performance and attraction, since the new generations are increasingly connected with technology [1].

It is considered that virtual reality in conjunction with 3D models not only allows a dynamic interaction, but also a sense of immersion. Thus, it is possible to greatly enhance all engineering concepts, where spatial skills are of great importance [2]. Also, 3D models generate a sense of immersion in students because they facilitate direct interaction with objects, improving spatial orientation skills in users, important concepts in areas such as engineering or medicine. In other words, the development of spatial orientation concepts (such as spatial positioning) is possible because of the immersion provided by the manipulation of three-dimensional models.

Also, 3D models are considered to allow the analysis and visualization of the smallest details of figures and spaces, in such a way that educational software produces an improvement in perspectives [3]. According to [4] cited by [5] a person manipulates a physically solid 3D object and rotates it, the rotations made in the hand are so fast, unconscious, and inaccurate that a formal reflection of such actions can hardly be made. However, in 3D software it is possible to constrain the direction of rotation and forces students to envision various strategies with respect to the motion and anticipate the final result of the transformation.

In short, having a 3D object allows for accurate and precise manipulation of the object. In addition, by continuously manipulating the 3D model, spatial orientation is developed, so that it is possible to determine the final position of the 3D model after a rotation before it is made. The latter is possible because the individual's brain quickly finds a reference point and forms a model of the object in his brain and mentally rotates it using the reference point.

Similarly, a study conducted by mathematics teachers in Colombia [5] on the effectiveness of a computer environment in the development of spatial skills yielded positive results. This consisted in the creation of a software for the representation of three-dimensional objects, described as follows: The first application elaborated for the development of spatial visualization is the software called cubes and cubes which corresponds to a micro world for the teaching and learning of spatial geometry by means of computational technologies. This educational software offers the possibility of

exploring space and three-dimensional objects in a novel and totally interactive way. Cubes and cubes allows the development of spatial visualization, perspective management and the ability to calculate volumes of irregular solids.

A clear example where the use of spatial skills is put into practice is the software activity called: "build a solid given the views", which aims to build a solid with the cubes, based on information from the top, side and front views of the result to be obtained.

This work is focused on developing spatial skills, which was achieved through the implementation of a software (cubes and cubes), which presented activities in which the students needed to locate reference points in order to carry out the constructions, for which they were given stimuli (profiles of the object to be built); in this way, the brain is trained to locate reference points and locate themselves in the environment easily.

On the other hand, in 2013 a study [6] (p.598) entitled: Three-Dimensional Sinus Imaging as an Adjunct to Two-Dimensional Imaging to Accelerate Education and Improve Spatial Orientation was developed, where three-dimensional images were implemented to two-dimensional images related to the study of otorhinolaryngology to improve the spatial orientation of students during a surgical process; this study concluded satisfactorily. The results of this study indicate that the addition of a 3-D educational module to traditional 2-D training significantly improves understanding of the anatomy and spatial orientation of the paranasal sinuses and surrounding structures. This finding adds to the growing body of literature supporting 3-D modeling and simulation as a positive contribution to education.

The three-dimensional images complemented the two-dimensional images by improving the spatial orientation of the students in relation to the paranasal sinuses. This was achieved because, by manipulating the three-dimensional images, the student was able to explore in more detail the concepts offered by the 2D image; thus, their spatial orientation was improved. Likewise, the improvement is evidenced in the results of a test, where the students who manipulated 3D images obtained better grades than the control group that only used images.

II. STATE OF THE ART

A. 3D Modeling

3D modeling is the process of constructing an object represented by a collection of points in three-dimensional space. "A 3D prototype requires two components: modeling and texturing. Using software, such as AutoCAD MAYA and Google SketchUp, any physical object can be modeled [7] (p. 8). Furthermore, 3D modeling is defined as the representation of objects in three dimensions (X, Y and Z). More specifically, modeling, is a process of creating a mathematical representation of surfaces using geometry [8]. Three-dimensional modeling is the process that allows the virtual formation of a physical object using software that has as its basis the three-dimensional coordinate system (X, Y, Z), in which points are placed "strategically" located that will be part of the 3D model of the desired physical object.

B. Three-dimensional Coordinate System

It is considered as, "a reference system formed by three straight lines or coordinate axes that intersect at a point called origin and a unit of measurement, these three lines are also called Cartesian coordinates, with respect to the XYZ system" [9] (p.6). A 3D image is modeled by means of an XYZ coordinate system (which starts from an origin), since this system, in addition to providing a height and width, allows to give it a depth. Consequently, this reference system becomes the basis of the three-dimensional image.

C. 3D graphics

Three-dimensional (3D) computer graphics, similar to two-dimensional or vector graphics, is a branch of computer-aided visualization. Their distinction is the ability to view three-dimensional (depth) data for the following conversion into two-dimensional static images or dynamic videos [10] (p.1). Three-dimensional images may resemble two-dimensional images, but the former have a unique feature, which is depth. Therefore, this feature gives the model the quality of being able to be rotated in space about its respective axis.

D. Spatial Location and Intuitive Trajectory

According to Newcombe & Huttenlocher cited in [11] spatial location and intuitive trajectory can be understood as the development of mental evocations that involve elaborating two reference systems: the one based on internal cues and the one based on external cues. Both systems are constructed from the point of view of personal position (p.123). Likewise, there are two types of cues that aid spatial orientation. The first type of cues is symbolic in nature need to be semantically processed, moreover, they are usually arrows or other symbolic cues such as direction words that refer to spatial locations. While the other signals are exogenous, i.e., of explicit character and appear peripherally in the same place [12].

Both authors agree that, in order to be able to locate oneself in a space, it is necessary to build reference systems, which are divided into two types, implicit and explicit. They conclude that these need to be processed by the brain in order to reach a correct spatial location.

E. Spatial Perception

The ability to be able to recreate the image of an object and manipulate it mentally has a significant practical application in fields such as mathematics, physics, architecture or engineering. This ability, known as Spatial Perception, is the most important of all those that an individual must possess for the practice of engineering [2] (p. 2).

In addition, spatial perception is an ability that allows people to recreate and manipulate an object mentally. Likewise, this is achieved by considering reference points of the object to be recreated.

F. Spatial Organization

According to [11] this concept refers to the development of spatial perspective and spatial trajectories in non-close environments. The development of spatial perspective consists of the construction of conical reference systems using reference points extensive to the person, with which he/she can locate and locate objects or places (p. 123). Likewise, [13] mentions

that: The spatial organization of the child evolves from: an egocentric location, in which the child does not distinguish the space occupied by his body, with that occupied by the objects around him and an objective location in which the child is able to discriminate the space occupied by his body and by each object (p. 7).

In short, spatial organization is an ability that allows people to locate themselves in a certain space, through the construction of reference systems based on egocentric location, so that the individual is located as the center of a certain space, around which objects are found.

G. Spatial Structuring

This concept refers to the understanding of spatial relationships that are represented using Euclidean or polar coordinates in two-dimensional or three-dimensional planes, which can represent locations or trajectories of objects in certain points of the plane or space [11] (p. 124). That is, spatial structuring consists of situating an object or body in function with a given perspective, as are the chordates; in this concept the individual does not refer to his own body in this same one.

H. Space Channels

According to [13], the finding of dependent spatial channels in the visual search task raises some general questions about the visual search paradigm in particular and visual processing in general. As mentioned in the introduction, the visual search paradigm seems to be one in which the subject is encouraged to treat spatial channels as independent since the spatial (see Table I).

TABLE I. SPATIAL CHANNELS. ADAPTED FROM [13]

Table with 2 columns: Channel Name and Description. Rows include Visual, Kinesiological, Touch, Auditory, Memory, and Labyrinth.

III. METHODOLOGY

A. Objective of the Research

To determine the extent to which 3D graphics improve spatial orientation skills in secondary school students.

B. Research Hypothesis

The use of 3D graphics improves orientation skills in secondary school students.

C. Sample Population

The population is constituted by 300 students of which 25 students were chosen through a simple random sampling, where the choice is random because it is desired to avoid errors

when choosing students who have ample spatial skills. Four sessions of fifty minutes each have been developed in the subject of Educational Technology, in which three-dimensional models will be used, in order to determine if they are useful for the development of spatial skills.

D. Data Collection Techniques and Instruments

The main data collection instrument is the psychometric test of spatial reasoning standardized by [14] to obtain reliable results. This test is structured as shown in Table II.

TABLE II. PSYCHOMETRIC TEST OF SPATIAL REASONING

Table with 2 columns: Criteria and Score. Lists five criteria for spatial reasoning and a total score of 40.

IV. PROCEDURE

A psychometric pre-test was taken in order to determine how developed the spatial skills of the project participants are. For the visualization, students are asked to run the 3D Builder program, then the acquired 3D models are shared with them and they are given a certain amount of time to manipulate them, then the students are asked to place the object in a certain position and determine a final position, they are also asked to draw the object from different perspectives. Then, the psychometric post-test of spatial reasoning is applied, with the objective of analyzing the effectiveness of the images for the development of spatial skills. For the data analysis, data will be collected from the grades obtained by the students.

V. DATA ANALYSIS AND INTERPRETATION

A. Data Analysis by Criteria

Fig. 1 shows that in a pre-test no participant achieved a maximum score (8 points); however, the best score was 7, which represents 12% of the sample; and a minimum score of 2. In addition, only 56% of the sample exceeded 4 correctly marked questions.

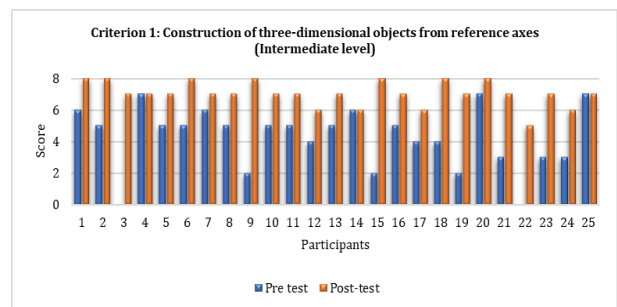


Fig. 1. Construction of Three-Dimensional Objects from Reference Axes (Intermediate Level).

From these data it can be inferred that, at the beginning, half of the students do not have developed the ability to construct three-dimensional objects, which is achieved through spatial orientation, since reference points are considered, and the image is formed in the mind.

Fig. 2 in the case of the second criterion, presents exercises similar to the previous one, but with a higher difficulty, it is evident that, in the pre-test, only 26% of the participants exceeded 50% of correct answers to the questions, due to the difficulty of the exercises. This is understandable considering that, as the difficulty of the exercises increases, the number of correct answers of the participants is reduced; demonstrating once again that spatial orientation skills were not previously fully developed. Therefore, in the post-test, although many exceed half of the correct answers, the scores are not high compared to the previous criterion.

According to Fig. 3, in the pre-test only 28% of the sample exceeded half of the answers marked correctly, in addition to the fact that there are marks that are too low. This shows a very low development of the criterion of mental rotation of objects.

The low scores are due to the fact that this criterion is not constantly worked on, given that in daily life it is rarely encountered [15].

Subsequently, in the post-test an increase in the ability of the participants was observed, going from a percentage of 26% to 80% of students who correctly marked more than half of the questions of criterion 3. This was achieved through the manipulation of three-dimensional objects and mental exercises, such as predicting the final position of the object after a turn [16] (p. 2).

Fig. 4 shows that the participants still have low scores in the pre-test, since only 38% marked correctly more than half of the questions.

However, compared to the previous criterion, which is relatively easier, in this case there is a higher percentage of students who scored more than half of the questions correctly. This is due to the fact that the previous exercises could have served as a previous practice that allowed them to improve in these scores.

Subsequently, there is evidence of progress in the students' grade point average, which is obtained through the manipulation of three-dimensional objects throughout the sessions.

Fig. 5 shows that in the pre-test 20% of the sample marked correctly more than half of the questions, demonstrating a deficiency in the criterion of mental deconstruction of three-dimensional objects to two-dimensional planes. This is due to the fact that this skill is rarely put into practice, so that there is no mastery of it. Likewise, after the exercises and in the post-test, in many cases there is no evidence of a considerable improvement in the results, since, in order to completely improve this skill, many more sessions and constant practice would be required.

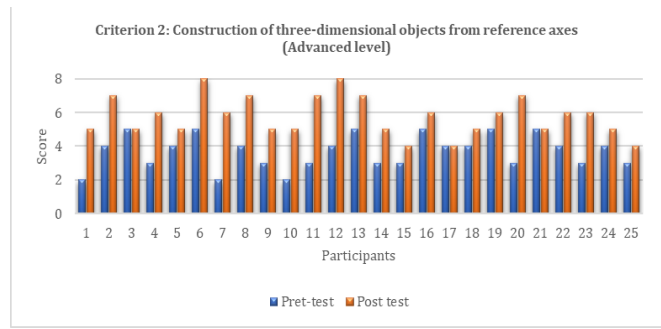


Fig. 2. Criterion 2: Construction of Three-dimensional Objects (Advanced Level).

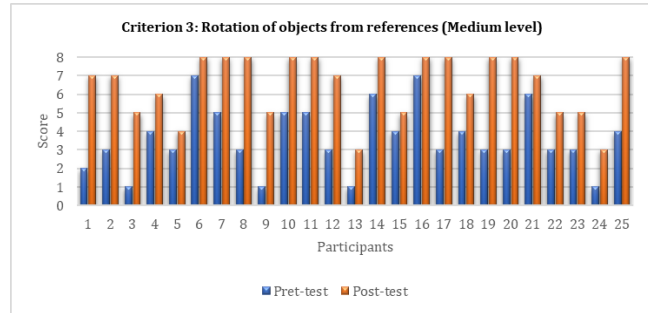


Fig. 3. Criterion 3: Rotation of Objects from References (Medium Level).

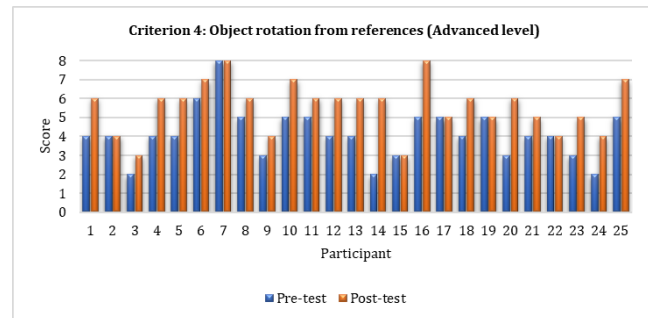


Fig. 4. Criterion 4: Rotation of Objects from References (Advanced Level).

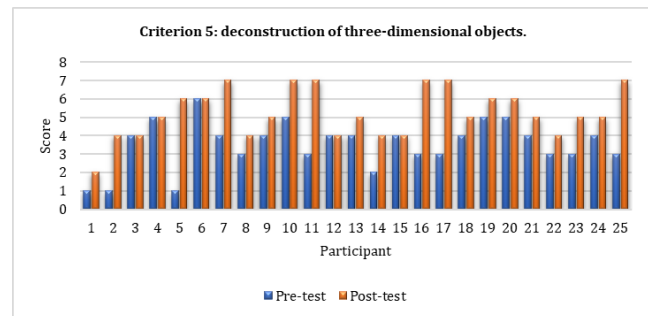


Fig. 5. Criterion: Deconstructions of Three-dimensional Objects.

B. Global Data Analysis

In Table III, in the global analysis of the results of both the pre-test and post-test, significant differences can be seen in both results, which we will analyze through the Histogram and Polygon of Frequencies.

TABLE III. TOTAL SCORES (OUT OF 40 POINTS)

Participants	Pre-test	Post-test
1	15	28
2	17	30
3	12	24
4	23	30
5	17	28
6	29	37
7	25	36
8	20	32
9	13	27
10	22	34
11	21	35
12	19	31
13	19	28
14	19	29
15	16	24
16	25	36
17	19	30
18	20	30
19	20	32
20	21	35
21	22	29
22	14	24
23	15	26
24	14	23
25	22	33

According to Fig. 6 in the polygon, 40% of the participants have grades that exceed half, while the lowest grade is 12/40, demonstrating that although there are students who have developed spatial skills; there is also a percentage of 48% who do not have very developed these skills [17].

Table IV shows that 8 students obtained a score between 29 and 32 points in the post-test, which represents 32% of a total of 25 students.

According to Fig. 7, it is evident that the minimum score was 23 out of a total of 40, while the highest percentage of scores belongs to the interval [29-32[which denotes an improvement in the student's spatial location skills. Since the average improved by 10 more correct scores.

Fig. 8 shows that there is a clear increase in the number of questions answered correctly. Likewise, the Post-Test graph is similar in shape to the Pre-Test graph. This would be due to the fact that the students were able to develop their spatial skills simultaneously, highlighting that they received the learning almost evenly, since most of them showed similar progress, based on the pre-test.

According to the results observed in Tables 5 and 6, the researcher's hypothesis has been validated, which mentions that

the use of images and three-dimensional models allows improving the development of orientation skills in students. This is due to the fact that the P value is less than 0.05, so that the null hypothesis is rejected. Thus, it is concluded that the hypothesis proposed in this work is correct.

Studies have reported that spatial capabilities can be used to effectively predict the performance of individuals in performing complex operations [18].

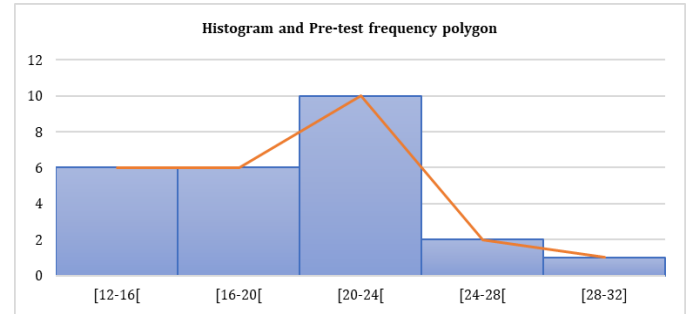


Fig. 6. Histogram and Polygon of Frequencies Pre-test.

TABLE IV. POST-TEST FREQUENCY TABLE

POST-TEST FREQUENCY TABLE				
Notes	f	F	fi	Hi
[23-26[4	4	0,16	0,16
[26-29[5	9	0,2	0,36
[29-32[8	17	0,32	0,68
[32-35[3	20	0,12	0,80
[35-38]	5	25	0,2	1,00
N	25		1.00	

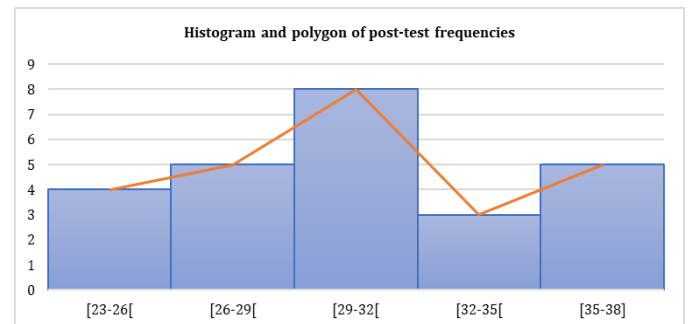


Fig. 7. Histogram and Polygon of Frequencies Post-test.

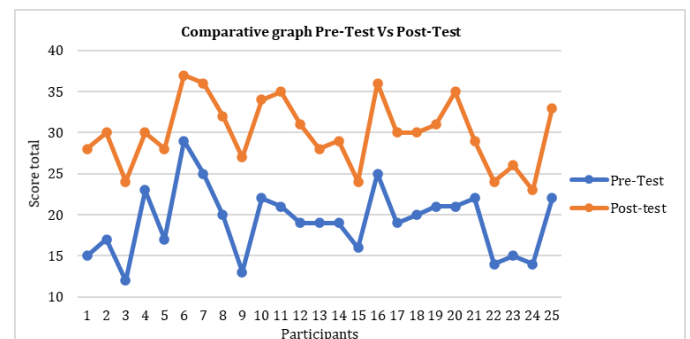


Fig. 8. Comparison between Pre-test and Post-test.

TABLE V. VALIDATION OF STUDENT'S T HYPOTHESIS

Steps	Procedure
Hypothesis formulation	In this part, two hypotheses are presented: the null hypothesis (H0) and the researcher's hypothesis (H1). H0=The use of 3D graphics does not improve the development of orientation skills in students. H1=The use of 3D graphics improves the development of orientation skills in students.
Significance level	It is a numerical value that is defined as the probability of making the decision to reject the null hypothesis. This value is 5% or 0.05%.
Choice of statistical test	In this case we have the Pre-test (Variable 1) and the Post-test (Variable 2).
p-value estimation	The Excel data analysis tool is used in which the T-test option for paired two-sample means is selected.
Decision making	If $p < 0.05$ then the Null hypothesis is rejected and the researcher hypothesis is accepted. If $p > 0.05$, the null hypothesis is accepted and the research hypothesis is rejected.

TABLE VI. P-VALUE ESTIMATION

	Variable 1	Variable 2
Media	19.1600	30.04
Variance	17.0567	16.70666667
Pearson correlation coefficient	0.8758	-
Hypothetical difference of means	0.0000	-
Degrees of freedom	24.0000	-
Statistic t	-26.5656	-
P(T<=t) one tail	0.0000	-
Critical value of t (one-tailed)	1.7109	-
P(T<=t) two-tailed	0.0000	-
Critical value of t (two-tailed)	2.0639	-

VI. FUTURE RESEARCH SCOPES

The future scope of the research involves experimentation with a larger population using more indicators with a qualitative approach where behaviors, qualities and attitudes can be observed in order to obtain reliable data. According to [19] visualization, as both the product and the process of creation, interpretation and reflection upon pictures and images, is gaining increased visibility in mathematics and mathematics education.

Currently the need to provide quality education to future generations has led to the development of new teaching methodologies; within this fact the tools provided by information technologies have been positioned as the future of learning [20].

VII. DISCUSSION AND CONCLUSIONS

The use of 3D graphics in the teaching-learning processes allows improving the spatial orientation skills of the students to a great extent. This result is evidenced by the increase in the

total scores obtained in the Post-test compared to the results of the Pre-test. Also, specifically, it was observed, on average, an increase from 47.9% to 75.1% of questions answered correctly, which in conjunction with the Student's t-test that gave a P value less than 0.05 demonstrating the reliability of the study and therefore the significant improvement of spatial orientation skills in students. On the other hand, criterion 2 shows the greatest increase in the students' scores, because the manipulation and rotation of objects was an activity that was performed in all learning sessions.

3D graphics can be used in several educational areas, in order to improve people's spatial skills. Likewise, it is considered that in order to improve spatial skills it is necessary to practice constantly, since it is necessary to put spatial skills into practice, taking into account that they are part of the multiple intelligences of people, specifically in spatial intelligence. In addition, spatial skills are very important in the training of professionals in areas such as engineering or architecture, which is why many universities take exams in relation to this type of skills.

Likewise, in order for students to efficiently develop spatial skills, such as spatial orientation, it is necessary for them to have a clear understanding of digital citizenship, so that they can handle technology in an appropriate and responsible manner. In addition, proper supervision by teachers must be carried out.

ACKNOWLEDGMENT

The research was made possible thanks to the support of the Universidad Nacional de San Agustín de Arequipa through UNSA INVESTIGA.

REFERENCES

- [1] Maraza-Quispe, B., Alejandro-Oviedo, O., Choquehuanca-Quispe, W. (2020). Towards a Standardization of Learning Behavior Indicators in Virtual Environments. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11. From https://thesai.org/Downloads/Volume11No11/Paper_19-Towards_a_Standardization_of_Learning_Behavior.pdf.
- [2] Suárez Quiróz, J., Rubio García, R., Gallegos Santos, R., & Gonzáles, M. (2004). Development of a trainer for spatial perception based on virtual reality using public domain technologies. Sustainability 2018, 10, 1074; doi:10.3390/su10041074.
- [3] Nurbekova, Z., Grinshkun, V., Aimicheva, G., Nurbekov, B. & Tuenbaeva, K. (2020). Project-Based Learning Approach for Teaching Mobile Application Development Using Visualization Technology. International Journal of Emerging Technologies in Learning (iJET), 15(8), 130-143. Kassel, Germany: International Journal of Emerging Technology in Learning. from <https://www.learnlib.org/p/217072/>.
- [4] Gutierrez, A., Pegg, J., & Lawrie, C. (2004). Characterization of Students' Reasoning and Proof Abilities in 3-Dimensional Geometry. ERIC. Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education, Vol 2 pp 511–518.
- [5] Hoyos Salcedo, E. A., & Acosta Minoli, C. A. (2014). Improvement of spatial visualization skills through the use of a computer environment. XV Virtual Educa Peru Meeting.
- [6] Yao WC, Regone RM, Huyhn N, Butler EB, Takashima M. (2014). Three-dimensional sinus imaging as an adjunct to two-dimensional imaging to accelerate education and improve spatial orientation. Laryngoscope. Mar;124(3):596-601. doi: 10.1002/lary.24316. Epub 2013 Oct 2. PMID: 23881572.
- [7] Torres, C. E., & Rodríguez Carrilo, J. (2019). Immersive learning environments and teaching to cyber generations. vol.45, e187369. ISSN 1678-4634. <https://doi.org/10.1590/s1678-4634201945187369>.

- [8] Ortega, A. J. (2016). Digital Fabrication: Introduction to 3D modeling and printing. Ministerio de Educacion, Cultura y Deporte. Colección Aula Mentor. <https://sede.educacion.gob.es/publiventa>.
- [9] Charro Arévalo, C., & Valencia Armijos, V. (2007). Three-dimensional model of the geological history of Cotipaxi Volcano. Ecuador.
- [10] Pytlík, M., & Kostolányová, K. (2018). 3D technologies in education. AIP Conference Proceedings. doi:<https://doi.org/10.1063/1.5079085>.
- [11] Zapateiro Segura, J. C., Poloche Arango, S. K., & Camargo Uribe, L. (2017). Spatial orientation: a teaching and learning path focused on locations and trajectories. TED, 18.
- [12] Ouellet, M., Santiago, J., Funes, M., & Lupiáñez, J. (1999). Spatial orientation of attention using temporal concepts. 36(1), 17–24. <https://doi.org/10.1037/a0017176>.
- [13] Pollatsek, A. and Digma, L. (2020) Dependent spatial channels in visual processing, *Cognitive Psychology*, Volume 9, Issue 3, Pages 326-352, ISSN 0010-0285, [https://doi.org/10.1016/0010-0285\(77\)90011-1](https://doi.org/10.1016/0010-0285(77)90011-1).
- [14] Maraza-Quispe, B., Alejandro-Oviedo, O., Fernández-Gambarini, W., Cisneros-Chavez, B., & Choquehuanca-Quispe, W. (2020). Analysis of YouTube as a tool for documentary research in higher education students. *Publications*, 50(2), 133-147. doi:10.30827/publications.v50i2.13949.
- [15] Strong, S., & Smith, R. (2001). Spatial visualization: fundamentals and trends in engineering graphics. *Journal of Industrial Technology*. The Official Electronic Publication of the National Association of Industrial Technology. Volume 18, Number 1 • November 2001 to January 2002.
- [16] Navarro Rosa., Saorín, José Contero Manuel, Piquer and Ana Conesa, Julián. (2004). el desarrollo de las habilidades de visión espacial y croquis en la ingeniería de producto.
- [17] Dastoli, C. (2018). Design, digital fabrication & 3d printing: a crash course for design students. Conference: 10th International Conference on Education and New Learning Technologies. DOI: 10.21125/edulearn.2018.1186.
- [18] Liao, KH. (2017). Las habilidades para comprender las relaciones espaciales, la orientación espacial y la visualización espacial afectan el rendimiento del diseño de productos 3D: utilizando el diseño de cajas de cartón como ejemplo. *Int J Technol Des Educ* 27, 131–147. <https://doi.org/10.1007/s10798-015-9330-3>.
- [19] Arcavi, A. (2003). El papel de las representaciones visuales en el aprendizaje de las matemáticas. *Edu. Semental. Matemáticas*. 52 , 215–241. <https://doi.org/10.1023/A:1024312321077>.
- [20] Maraza-Quispe, B., Sotelo-Jump, A., Alejandro-Oviedo, O., Quispe-Flores, L., Cari-Mogrovejo, L., Fernandez-Gambarini, W and Cuadros-Paz, L. (2021). “Towards the Development of Computational Thinking and Mathematical Logic through Scratch” *International Journal of Advanced Computer Science and Applications*, 12(2), <http://dx.doi.org/10.14569/IJACSA.2021.0120242>.

Symptoms-based Fuzzy-Logic Approach for COVID-19 Diagnosis

Maad Shatnawi¹

Department of Electrical Engineering Technology
Higher Colleges of Technology
Abu Dhabi, UAE

Anas Shatnawi²

Berger-Levrault
Montpellier, France

Zakarea AlShara³

Department of Software Engineering
Jordan University of Science and Technology
Irbid, Jordan

Ghaith Husari⁴

Department of Computer Science
East Tennessee State University
Johnson City, Tennessee

Abstract—The coronavirus (COVID-19) pandemic has caused severe adverse effects on the human life and the global economy affecting all communities and individuals due to its rapid spreading, increase in the number of affected cases and creating severe health issues and death cases worldwide. Since no particular treatment has been acknowledged so far for this disease, prompt detection of COVID-19 is essential to control and halt its chain. In this paper, we introduce an intelligent fuzzy inference system for the primary diagnosis of COVID-19. The system infers the likelihood level of COVID-19 infection based on the symptoms that appear on the patient. This proposed inference system can assist physicians in identifying the disease and help individuals to perform self-diagnosis on their own cases.

Keywords—COVID-19; coronavirus diagnosis; fuzzy inference system; fuzzy logic; fuzzy rules; expert systems

I. INTRODUCTION

The Coronavirus disease 2019 (COVID-19) pandemic has seriously affected all aspects of our life including health, education, economy, travel, and entertainment. Spreading rapidly across borders, the coronavirus disease has created a global health crisis and caused numerous death cases all over the world. Coronaviruses are a wide group of viruses that cause sickness starting from the common cold and up to very severe infections leading to death in several situations [1-3]. The common COVID-19 symptoms that are normally seen within 2 to 14 days are cold, dry cough, fever, flu, breathing difficulties, throat sore, and headache [4-8].

Since no therapeutic drug has been confirmed for COVID-19 till this date, the early diagnosis and preventions are essential to control and break down the chain of COVID-19 by immediate isolation of the infected person from the healthy population [9] [10]. The most common methods that global healthcare systems are currently using for Covid-19 identification are Real-Time Polymerize Chain Reaction (RT-PCR) tests in addition to chest Computerized Tomography (CT) scan and X-ray imaging. However, PCR testing requires several hours to get the results and suffers from high false positive rates and false negative rates which means it does not

identify all infections, and therefore, PCR should not be used as the only criterion for detecting COVID-19 patients [11-13].

A number of studies reported that chest CT scan has considerably higher COVID-19 diagnosis sensitivity than RT-PCR. On the other hand, CT scans and X-rays have the following limitations. First, CT scans have high false negative rates, as they are unable to distinguish coronary tissue from non-coronary tissue. A large number of COVID-19 patients have normal chest CTs or X-rays. Second, CT scans are unable to discriminate between cancerous tissue, cysts, and coronary tissue. Third, the nonappearance of an anomaly on either a chest X-ray or CT scan does not necessarily eliminate being COVID-19 infected. Fourth, chest CT scans and X-ray cannot precisely differentiate between COVID-19 and other respiratory infections such as seasonal flu. Fifth, CT scanning machines are complex equipment that should be carefully sanitized between potential COVID-19 patients, and there is a risk that the virus remains on the surfaces of CT scanning rooms. Finally, moving potential COVID-19 patients to and from a CT scanning room increases the hazard of spreading the virus within healthcare centers [13-18].

As a result, integrating these methods with a symptom-based diagnosis method will lead to more accurate identification results. In this work, we propose a smart fuzzy inference system to diagnose the COVID-19 based on the symptoms that appear on the patient.

The rest of the paper is organized as follows. Section II presents a background of the fuzzy inference systems and its applications in medical diagnosis. In Section III, we describe the design of our COVID-19 inference system. We evaluate the effectiveness of our approach in Section IV. Section V presents concluding remarks and highlights future directions.

II. FUZZY INFERENCE SYSTEMS: BACKGROUND AND RELATED WORK

Fuzzy Inference Systems (FIS) use fuzzy reasoning in order to represent the knowledge of experts about certain problems in human-like decision-making. These systems are based on fuzzy logic modeling and allow attaining solutions based on

linguistic terms. They are principally useful in cases where human knowledge is available but there is no sufficient information to feed traditional mathematical model variables [19-25]. The fuzzy inference system is made up of four main modules; fuzzification module, knowledge base, inference engine, and defuzzification module as shown in Fig. 1.

The most commonly used fuzzy inference technique is the Mamdani model [26, 27]. The Mamdani fuzzy inference process is performed in four consequent stages; fuzzification, rule evaluation, rule output aggregation, and defuzzification. The fuzzification module maps the crisp input value into a degree of membership of fuzzy sets by applying fuzzification membership functions. A membership function returns a value between zero (for non-membership) and one (for full-membership). The Knowledge base includes the IF-THEN rules that are provided by field experts. The rules are in the form [28]:

IF (A is x) AND (B is y) AND (C is z) ... THEN (R is m) (1)

where A, B, and C represent the input variables while x, y, and z represent the corresponding linguistic terms (e.g., yes, no), R represents the rule output variable and m represent the corresponding linguistic term (e.g., high risk, medium risk, low risk). The defuzzification module converts the output of the inference engine into a crisp output value. The Centroid or the Center of Gravity (COG) method is the most popular defuzzification technique where the weighted average of the area bounded by the aggregated membership function curve of the output variable is considered the crisp output value [13, 29, 30].

The final defuzzified output value using the centroid method is calculated by the following equation:

$$d_{COG}(M) = \frac{\int_{-M}^M M(z)z dz}{\int_{-M}^M M(z) dz} \quad (2)$$

where M represents the membership function of the output variable.

Fuzzy inference systems have been widely used in the medical diagnosis of different diseases. Lee and Wang [31] presented a fuzzy expert system based on fuzzy ontology as a decision support model for diabetes. Mayilvaganan and Rajeswari [32] proposed high blood pressure fuzzy logic classifier. Ekong et al. [33], Djam et al. [34] and Sharma et al. [35] proposed fuzzy expert systems for malaria diagnosis. Chandra [36] suggested a fuzzy expert system for migraine analysis and diagnosis.

Several fuzzy inference models have been proposed for heart disease diagnosis such as Kumar [37], Adeli and Neshat [38], Kumar and Kaur [39], Kasbe and Pippal [40], Allahverdi et al. [41], Oad et al. [42] and Subbulakshmi et al. [43]. A number of fuzzy inference models were proposed for cancer detection [44] such as Keleş et al. [45], Balanica et al. [46] and Latha et al. [47] for breast cancer diagnosis, Lavanya et al. [48] for lung cancer diagnosis, and Saritas et al. [49] for prostate cancer diagnosis. Kolhe et al. [50] presented a fuzzy-logic based approach for disease-diagnosis in crops.

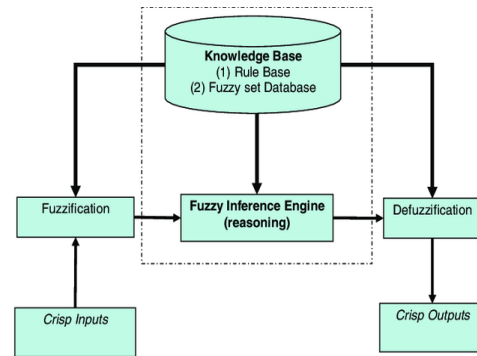


Fig. 1. The Fuzzy Inference System Block Diagram.

Patel et al. [51], Anand et al. [52], Mishra et al. [53] and Sundararaman et al. [54] proposed Asthma fuzzy diagnosis systems. Damirchi-Darasi et al. [55] proposed fuzzy rule-based expert system for the diagnosis of spinal cord disorders. Kadhim et al. [56] suggested a fuzzy expert system for back pain diagnosis. Zarei et al. [57] suggested a fuzzy modeling and control of HIV infection and [58] proposed diagnosis of HIV/AIDs using Fuzzy Cluster Means Algorithm.

Faisal et al. [59] employed an Adaptive Neuro-Fuzzy Inference System to predict the degree of risk of dengue patients. Saikia and Dutta [60] applied FIS to diagnosis of the Dengue disease. Alrashoud [61] proposed a Hierarchical Fuzzy Inference system for dengue fever diagnosis. Shaaban et al. [13] introduced a hybrid COVID-19 diagnosis system through fuzzy inference and deep neural networks based on four laboratory data which are White Blood Cell (WBC), Lymphocyte (LYM), Monocytes (MON), and Locate Dehydrogenase (LDH).

III. COVID-19 INFERENCE SYSTEM

In this work, a smart fuzzy inference system is proposed for the early detection of COVID-19 based on the patient symptoms including cold, cough, fever, flu, breathing difficulties, throat infection and headache [8]. The proposed system infers the likelihood level of COVID-19 infection based on the symptoms that appear on the patient. The COVID-19 fuzzy inference system is designed by identifying the input and output variables in addition to the fuzzy sets and membership functions of each variable. Afterward, a set of fuzzy rules that are connecting input variables with output variables are set. The proposed inference system is aims at diagnosing the COVID-19 based on the patient data.

We applied the Mamdani Fuzzy model to build the COVID-19 inference system. We define 9 symptoms as the input variables to the inference system. We group these variables into two categories; most common symptoms and less common symptoms. The most common symptoms category includes fever, tiredness, and dry cough while the less common symptoms category includes diarrhea, sore throat, headache, conjunctivitis, loss of taste or smell, and breathing difficulties. The output variable is risk of being COVID-19 infected. The COVID-19 inference system is illustrated in Fig. 2.

A. Membership Functions

Each input variable has two Gaussian membership functions as shown in Table I. The Gaussian membership function $\text{gaussmf}[\sigma \mu]$ is defined by its mean μ and standard deviation σ . The fever variable is represented by the body temperature which ranges between 36.5 and 42°C as presented in Fig. 3 while each of the remaining input variables has a level in the range from 0 to 5 as indicated in Fig. 4 [62, 63]. The output variable ranges from 0 to 100 and it has four Gaussian membership functions; low risk, medium risk, high risk and very high risk as presented in Fig. 5.

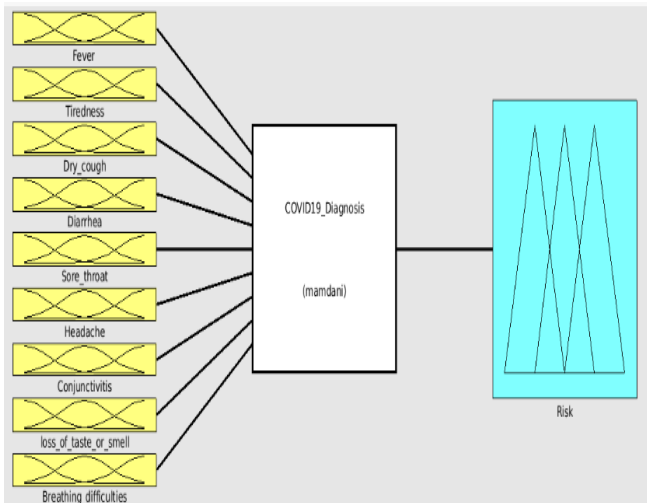


Fig. 2. COVID-19 Fuzzy Inference System.

TABLE I. MEMBERSHIP FUNCTIONS OF THE INPUT VARIABLES

Num ber	Variable name	Rang e	Membership Function 1 (No)	Membership Function 2 (Yes)
1	Fever	36.5- 42°C	'gaussmf',[0.5 36.5]	'gaussmf',[2 42]
2	Tiredness	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
3	Dry cough	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
4	Diarrhea	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
5	Sore throat	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
6	Headache	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
7	Conjunctiviti s	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
8	Loss of taste or smell	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]
9	Breathing difficulties	0-5	'gaussmf',[0.7 0]	'gaussmf',[1.85 5]

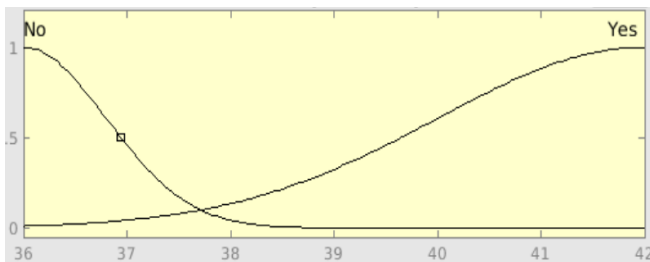


Fig. 3. The Fever Membership Functions.

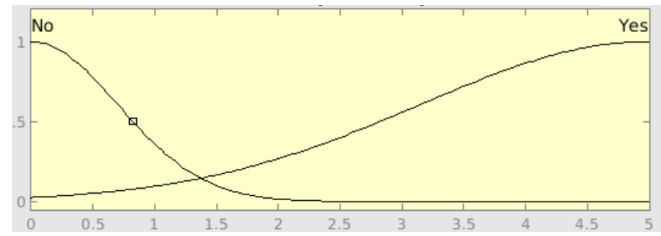


Fig. 4. The Tiredness Membership Functions.

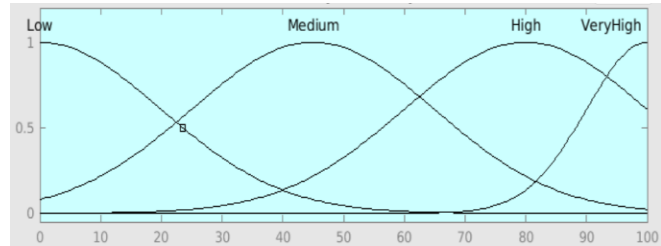


Fig. 5. The Risk of being COVID-19 Infected Membership Functions.

B. Fuzzy Rules

We define the following linguistic fuzzy rules:

- Rule 1: If 3 symptoms of Category-1 AND at least 4 symptoms of Category-2 are present → Very high risk of COVID-19 infection.
- Rule 2: If 3 symptoms of Category-1 AND less than 4 symptoms of Category-2 are present → High risk of COVID-19 infection.
- Rule 3: If the 3 symptoms of Category-1 are present → High risk of COVID-19 infection.
- Rule 4: If 2 symptoms of Category-1 AND at least 2 symptoms of Category-2 are present → High risk of COVID-19 infection.
- Rule 5: If 1 symptom of Category-1 AND at least 4 symptoms of Category-2 are present → High risk of COVID-19 infection.
- Rule 6: If 2 symptoms of Category-1 AND 1 symptom of Category-2 are present → Medium risk of COVID-19 infection.
- Rule 7: If 1 symptom of Category-1 AND 3 symptoms of Category-2 are present → Medium risk of COVID-19 infection.
- Rule 8: If 1 symptom of Category-1 AND 2 symptoms of Category-2 are present → Medium risk of COVID-19 infection.
- Rule 9: If no symptoms of Category-1 AND 6 symptoms of Category-2 are present → Medium risk of COVID-19 infection.
- Rule 10: If 2 symptoms of Category-1 AND no symptoms of Category-2 are present → Low risk of COVID-19 infection.

- Rule 11: If only 1 symptom of Category-1 AND 1 symptom of Category-2 are present → Low risk of COVID-19 infection.
- Rule 12: If only 1 symptom of Category-1 is present AND no symptoms of Category-2 → Low risk of COVID-19 infection.
- Rule 13: If no symptoms of Category-1 AND less than 6 symptoms of Category-2 are present → Low risk of COVID-19 infection.

C. Defuzzification of the Output

Based on the input patient symptoms, the inference system initiates a set of fuzzy rules where each rule produces an output. Fuzzy operator “min” was used for generating the output fuzzy set by taking every rule that satisfied the AND operational logic for a given set of input values. Then the output fuzzy set of each rule was combined into a single fuzzy set by the aggregation process. The single fuzzy set was defuzzified into a single numeric output value using the Centroid method to determine the percentage risk level of being COVID-19 infected.

IV. SYSTEM TESTING AND EVALUATION

This section presents the evaluation of our approach in the following terms: (a) system validation based on the feedback of field experts and (b) system testing using generated mock patient data.

A. System Validation

For the evaluation of the proposed system, we define two main research questions (RQs) and received feedback from the field experts in the healthcare domain using a survey. The RQs are:

- RQ1: Are the COVID-19 Symptoms considered by our approach correct? The goal of this research question is to evaluate the list of COVID-19 symptoms that are used to build our approach.
- RQ2: Are the fuzzy rules correct? The goal of this research question is to evaluate the correctness of the set of fuzzy rules that are used by our approach to decide whether a person is infected by COVID-19 or not.

1) Study design: To ease the accessibility to the survey, we created a web-based survey using Google forms¹. To test the relevance of the survey’s questions before publishing, we conducted a pilot with five candidate participants from the healthcare domain. Each tester practitioner evaluated all questions and their related answers. As a result, they propose minor revisions of the survey. The survey was prepared based on three main sections as follows:

- The first section aims to allow us to describe the participants of this survey by collecting general information about them such as their ages, levels of

experience in healthcare and medicine, professions, organizations and countries.

- The second section aims to evaluate the correctness of the set of symptoms related to COVID-19. To this end, each practitioner is asked to select a set (subset) of COVID-19 symptoms among the ones used in our approach. In order to identify COVID-19 symptoms that are not used by our approach, we make it also possible for a practitioner to add new COVID-19 symptoms.
- The last section includes questions related to the evaluation of the fuzzy rules defined in our approach. We asked the participants to evaluate each rule based on three options: Totally Agreed, Partially Agreed and Not Agreed. Totally agreed means that participants confirm our rule are correct following their experiences. Partially agreed refers to the case where participants agreed with this rule, but they do not consider it as correct in all cases. Meaning, the rules are correct for most cases, but not all compared to the COVID-19 patients based on their experiences. Not agreed means that participants do not agree with this rule. That means the rule should be modified.

To avoid prejudice, the survey was distributed to diverse participants from different health professions, levels of experience, organizations and countries. This distribution is based on social media and direct contact of health organizations such as hospitals and medical centers. One hundred participants have been invited to participate in the survey. They also have been requested to forward the survey to their networks.

2) Results

a) *Participants*: We have received 90 responses in total from participants from 11 different countries on four continents. The participants are also from different professions as presented in Fig. 6 where they cover almost all health domains that are related to COVID-19. Following their experience in the health domain, 58.9%, 28.8% and 12.2% of the participants have respectively more than 10 years, between 5 and 10 years and less than 4 years. The results show that 56.2%, 30%, 4.5% and 3.4% of the participants work respectively in Hospitals, Medical Centers, Universities and Pharmacies. As a result, the participants are diverse in their professions, type of health organizations, levels of experiences and geographical areas which means that they represent a good-enough sample that does not include prejudice in the answers to the survey questions.

b) *RQ1*: Are the COVID-19 Symptoms considered by our approach correct?

The results of the survey show that 90% of the participants agreed with us that these symptoms can be strongly used in COVID-19 diagnoses.

¹ This survey is accessible through:
https://docs.google.com/forms/d/e/1FAIpQLScDpa2uBCJg6366l-7vOtFjVd9ICet_nh8k4QXc5k0vZ2xv6Q/viewform

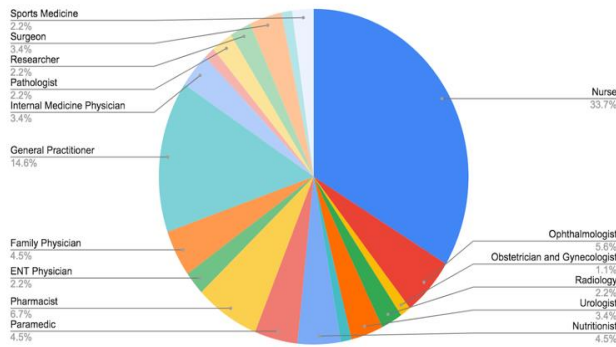


Fig. 6. Professions of the Survey Participants.

Fig. 7 shows the results of the evaluation of COVID-19 symptoms we used in our approach. The results show that the participants agreed with us for most of these symptoms. 6 of these symptoms (i.e., fever, breathing difficulties, loss of taste or smell, headache, dry cough, and tiredness) have been selected by more than 82% of the participants where fever is ranked as number one by 92.2% of the participants. The sore throat and diarrhea symptoms have been also selected by a quite number of participants. This means that the symptoms considered in our approach are representative compared to real COVID-19 cases based on the experience of the participants.

As it is allowed for the participants to add extra COVID-19 symptoms, we received only two extra symptoms: the stress and the body pain where each has been selected by one practitioner. These are rare symptoms. Thus, their absence will not negatively impact our approach.

Further, the results show that 85.5% of the participants agreed with us that Fever, Tiredness and Dry cough have more correlation with COVID-19 than the other Symptoms (from Question 8 in the survey). This confirms our decision to include these three symptoms in Category-1 that have higher weights in the fuzzy rules.

c) RQ2: Are the fuzzy rules correct?

Fig. 8 shows the results of evaluating the fuzzy rules. These results show that all of the rules are either partially or totally accepted by more than 80% of the participants. For example, Rule 1 has been accepted by 97.88% (77.88% + 20%) of the participants. This means that the fuzzy rules can be used to build the COVID-19 inference system.

B. System Testing

Based on a certain input of patient symptoms, the inference system initiates a set of fuzzy rules where each rule produces an output. Then, aggregation and defuzzification is performed to generate a single overall output through the process of

Centroid calculation. This final output represents the percentage risk of being COVID-19 infected. The proposed system is tested on some mock patient cases and the results are presented in Table II. Fig. 9 illustrates the rule evaluation process for a high-risk case while Fig. 10 and Fig. 11 illustrate the rule evaluation process for medium-risk and low-risk cases, respectively.

The 3D surface view for the rule that relates the COVID-19 infection risk to both fever and tiredness symptoms are demonstrated in Fig. 12. The dark blue surface represents a very low infection risk (less than 54%) when both symptoms are low. The green surface represents a higher risk (between 54% and 60%) when one of the two symptoms is high. The yellow surface represents a 65% risk when both symptoms are high. The 3D surface view for the rule that relates the COVID-19 infection risk to both breathing difficulties and sore throat symptoms is demonstrated in Fig. 13. Unfortunately, the 3D surface viewer can show the relationship of the output variable with only two input variables and since high infection risk (more than 65%) exists when at least 3 variables are high, this case cannot be viewed through this tool.

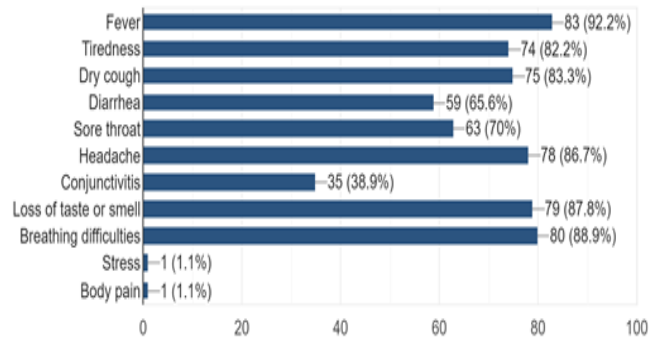


Fig. 7. The Results of Evaluating the Symptoms of COVID-19.

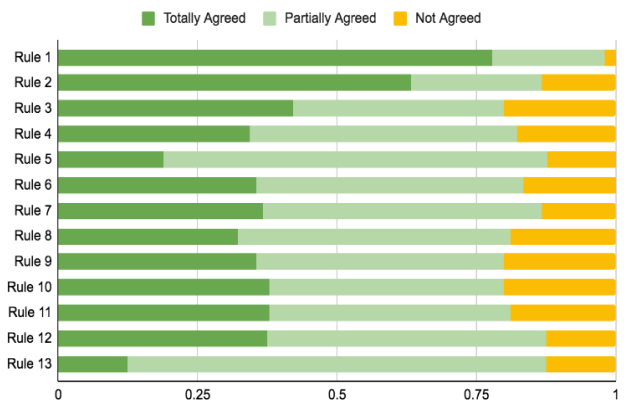


Fig. 8. Evaluation of Fuzzy Rules.

TABLE II. SYSTEM TESTING WITH POSSIBLE VALUES FOR PATIENT SYMPTOMS

Case	Fever (36-42°C)	Tiredness (0-5)	Dry cough (0-5)	Diarrhea (0-5)	Sore throat (0-5)	Headache (0-5)	Conjunctivitis (0-5)	Loss of taste or smell (0-5)	Breathing difficulties (0-5)	Output (Risk Percentage)	Risk level
1	38.5	5	5	5	5	5	5	5	5	84.84	High
2	42	5	5	5	5	5	5	5	5	92.33	Very high
3	39	3	3	3	3	3	3	3	3	73.35	Medium
4	41	5	2	1	3	3	3	0	1	67.52	Medium
5	37	1	0	1	0	3	3	5	4	51.50	Low
6	40	5	3	3	3	3	3	3	3	87.14	High
7	39.5	0	5	5	5	5	5	5	5	71.41	Medium
8	40	5	4	3	3	3	3	3	3	89.37	Very high
9	41	1	2	1	3	3	3	0	1	61.77	Medium
10	38	1	0	1	0	3	3	0	1	51.73	Low
11	40	5	4	1	3	3	3	0	1	73.22	Medium
12	37	1	1	3	4	5	5	5	4	54.52	Low
13	40	5	4	1	3	3	3	3	1	87.10	High

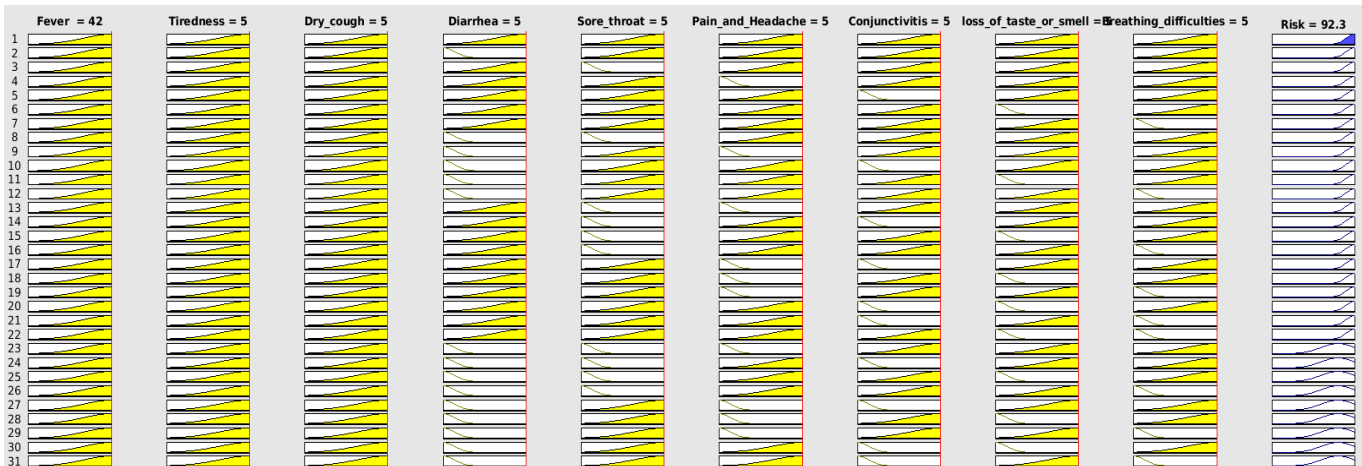


Fig. 9. Rule Evaluation for a High-Risk Case.

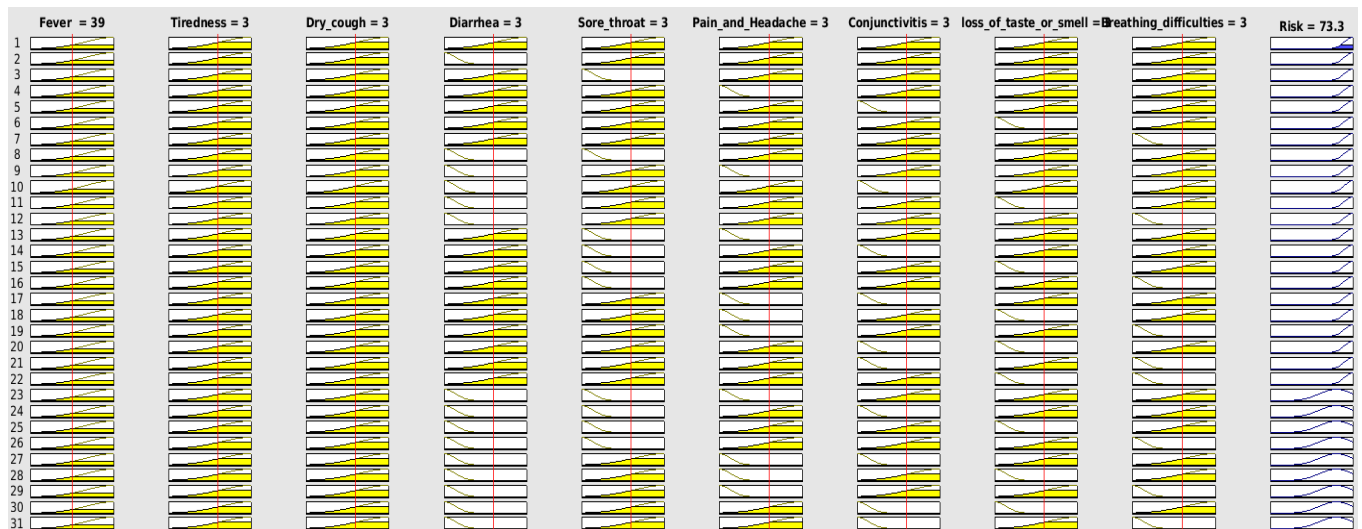


Fig. 10. Rule Evaluation for a Medium-Risk Case.

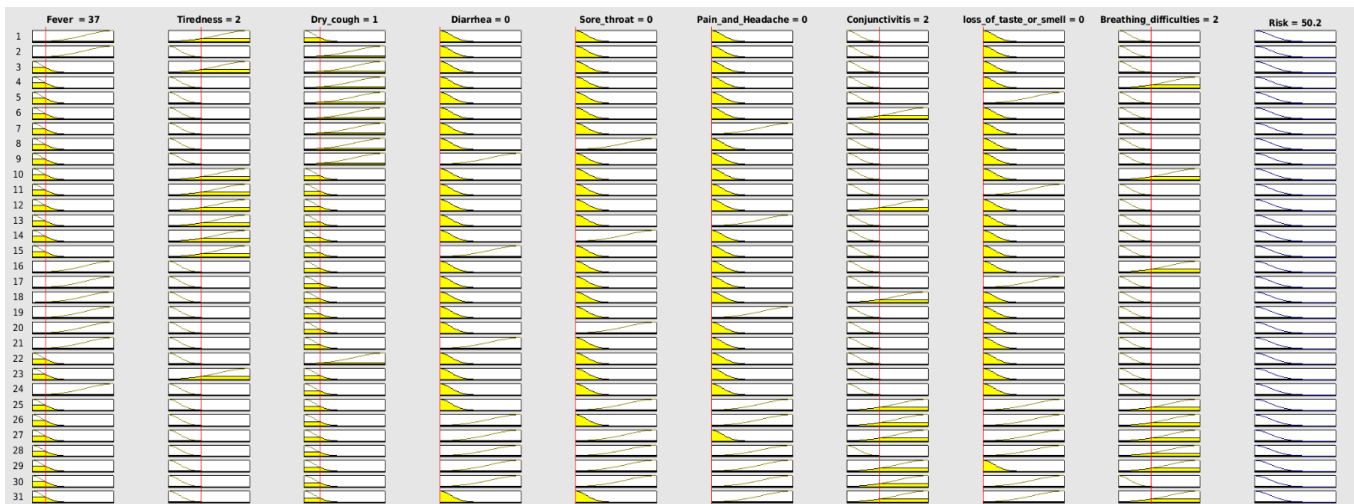


Fig. 11. Rule Evaluation for a Low-Risk Case.

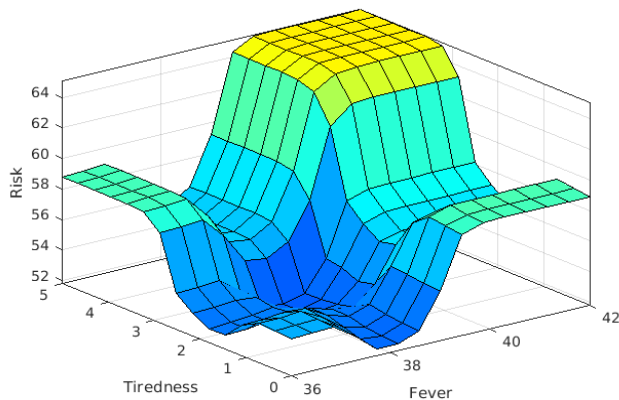


Fig. 12. COVID-19 Infection Risk vs. Fever and Tiredness.

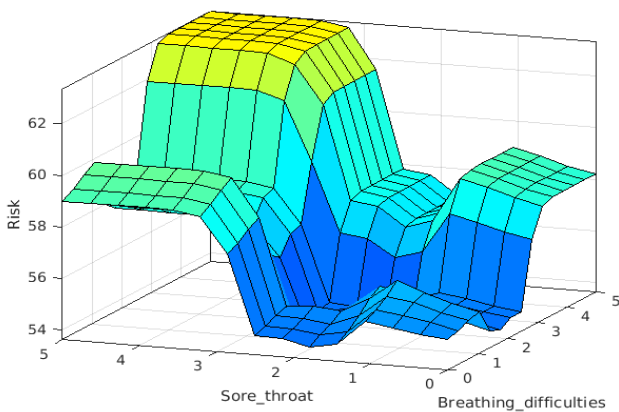


Fig. 13. COVID-19 Infection Risk vs. Breathing Difficulties and Sore Throat.

V. CONCLUSIONS AND FUTURE DIRECTIONS

We proposed a smart fuzzy inference system for the initial identification of COVID-19. The system infers the risk level of being COVID-19 infected based on the symptoms that appear on patients. The symptoms considered are fever, tiredness, and dry cough, diarrhea, sore throat, headache, conjunctivitis, loss of taste or smell, and breathing difficulties. This inference

system can assist physicians in identifying the disease. Although the proposed system cannot provide a very accurate COVID-19 identification, it can be integrated with other identification techniques such as PCR test and CT scan to work together to confirm infected cases.

In future work, we are planning to implement this diagnosis system into a web application to allow individuals to perform self-diagnosis on their own cases. The work can be extended to include other patient data such as blood pressure, breathing air peak-flow-rate, and having a chronic disease. One of the interesting future directions is to apply data mining techniques to generate fuzzy rules from patient data.

REFERENCES

- [1] Jamshidi, S. J. Zargaran, A. Talaei-Khoei and M. Kakavandi, "Modelling and Forecasting The Number of Confirmed Cases and Deaths from COVID-19 Pandemic in USA from April 12th to May 21st, 2020," medRxiv, 2020.
- [2] "WHO Coronavirus Disease (COVID-19) Dashboard," December 2019. [Online]. Available: <https://www.who.int/>. [Accessed 1 March 2021].
- [3] "Coronavirus Update," Worldometers, [Online]. Available: <https://www.worldometers.info/coronavirus/>. [Accessed 1 March 2021].
- [4] "Centers for Disease Control and Prevention," U.S. Department of Health & Human Services, 2020. [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/index.html>. [Accessed 1 March 2021].
- [5] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D. S. Hui and others, "Clinical characteristics of coronavirus disease 2019 in China," *New England journal of medicine*, vol. 382, no. 18, pp. 1708--1720, 2020.
- [6] H. Arslan and H. Arslan, "A new COVID-19 Detection Method from Human Genome Sequences using CpG island Features and KNN Classifier," *Engineering Science and Technology, an International Journal*, 2021.
- [7] "Journal of the American Medical Association," 2020. [Online]. Available: <https://www.ama-assn.org/about/publications-newsletters/jama-network>. [Accessed 1 March 2021].
- [8] N. Dhiman and M. Sharma, "Fuzzy Logic Inference System for Identification and Prevention of Coronavirus (COVID-19)," *International Journal of Innovative Technology and Exploring Engineering*, 2020.
- [9] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun and L. Xia, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, p. 200642, 2020.

- [10] "Emerging respiratory viruses, including COVID-19: methods for detection, prevention, response and control," World Health Organization, 2020. [Online]. Available: <https://openwho.org/courses/introduction-to-ncov>.
- [11] L. Fan and S. Liu, "CT and COVID-19: Chinese experience and recommendations concerning detection, staging and follow-up," *European Radiology*, vol. 30, no. 9, p. 5214–5216, 2020.
- [12] A. Tahamtan and A. Ardebili, "Real-time RT-PCR in COVID-19 detection: issues affecting the results," *Expert Review of Molecular Diagnostics*, vol. 20, no. 5, pp. 453-454, 2020.
- [13] W. M. Shaban, A. H. Rabie, A. I. Saleh and M. Abo-Elsoud, "Detecting COVID-19 patients based on fuzzy inference engine and Deep Neural Network," *Applied Soft Computing*, vol. 99, p. 106906, 2021.
- [14] Z. Ye, Y. W. Y. Zhang, Z. Huang and B. Song, "Chest CT manifestations of new coronavirus disease 2019 (COVID-19): a pictorial review," *European Radiology*, vol. 30, p. 4381–4389, 2020.
- [15] S. a. F. A. a. J. M. Inui, N. Kunishima, S. Watanabe, Y. Suzuki, S. Umeda and Y. Uwabe, "Chest CT Findings in Cases from the Cruise Ship Diamond Princess with Coronavirus Disease (COVID-19)," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, pp. 1-17, 2020.
- [16] R. Yang, X. Li, H. Liu, Y. Zhen, X. Zhang, Q. Xiong, Y. Luo, C. Gao and W. Zeng, "Chest CT severity score: an imaging tool for assessing severe COVID-19," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, pp. 1-23, 2020.
- [17] W. Xia, J. Shao, Y. Guo, X. Peng, Z. Li and D. Hu, "Clinical and CT features in pediatric patients with COVID-19 infection: Different points from adults," *Pediatric pulmonology*, vol. 55, no. 5, pp. 1169–1174, 2020.
- [18] H. Derakhshanfar, B. Sobouti, S. Fallah, Z. Soltantooyeh, S. Rahimi and S. Mirbaha, "A review on the effect of Coronavirus infections on respiratory problems in Children," *Journal of Critical Reviews*, vol. 7, no. 7, pp. 1-5, 2020.
- [19] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338--353, 1965.
- [20] L. A. Zadeh, "Fuzzy sets," in *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*, World Scientific, 1996, pp. 394--432.
- [21] T. Heske and J. N. Heske, *Fuzzy logic for real world design*, Annabooks, 1996.
- [22] M. F. Azeem, *Fuzzy inference system: theory and applications*, BoD--Books on Demand, 2012.
- [23] A. Tagarakis, S. Koundouras, E. Papageorgiou, Z. Dikopoulou, S. Fountas and T. Gemtos, "A fuzzy inference system to model grape quality in vineyards," *Precision Agriculture*, vol. 15, no. 5, pp. 555--578, 2014.
- [24] F. Chabni, R. Taleb, A. Benbouali and M. A. Bouthiba, "The application of fuzzy control in water tank level using Arduino," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, pp. 261-265, 2016.
- [25] V. D. B. Huynh, P. Nguyen, Q. Nguyen and P. T. Nguyen, "Application of fuzzy analytical hierarchy process based on geometric mean method to prioritize social capital network indicators," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 12, pp. 182-186, 2018.
- [26] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International journal of man-machine studies*, vol. 7, no. 1, pp. 1-13, 1975.
- [27] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE transactions on computers*, vol. 12, pp. 1182--1191, 1977.
- [28] Y. I. Daradkeh and I. Tvoroshenko, "Technologies for Making Reliable Decisions on a Variety of Effective Factors using Fuzzy Logic," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 43-50, 2020.
- [29] S. M. Heikel and M. E. Khedr, "Realization of strong ports and shipping services in developing countries using multi-criteria selection algorithm," in *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, 2010.
- [30] S. Naaz, A. Alam and R. Biswas, "Effect of different defuzzification methods in a fuzzy based load balancing application," *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 5, p. 261, 2011.
- [31] C.-S. Lee and M.-H. Wang, "A fuzzy expert system for diabetes decision support application," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 139--153, 2010.
- [32] M. Mayilvaganan and K. Rajeswari, "Human Blood Pressure Classification Analysis Using Fuzzy Logic Control System in Datamining," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 3, no. 1, pp. 306-306, 2014.
- [33] B. Ekong, I. Ifiok, I. Udoeka and J. Anamfiok, "Integrated Fuzzy based Decision Support System for the Management of Human Disease," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, 2020.
- [34] X. Djam, G. Wajiga, Y. Kimbi and N. Blamah, "A fuzzy expert system for the management of malaria," *International Journal of Pure and Applied Sciences and Technology*, 2011.
- [35] P. Sharma, D. Singh, M. K. Bandil and N. Mishra, "Decision support system for malaria and dengue disease diagnosis (DSSMD)," *International Journal of Information and Computation Technology*, vol. 3, no. 7, pp. 633--640, 2013.
- [36] V. Chandra, "Fuzzy expert system for migraine analysis and diagnosis," *International Journal of Science and Research*, vol. 3, no. 6, pp. 956--959, 2014.
- [37] A. Kumar, "Diagnosis of heart disease using Advanced Fuzzy resolution Mechanism," *International Journal of Science and Applied Information Technology*, vol. 2, no. 2, pp. 22--30, 2013.
- [38] A. Adeli and M. Neshat, "A fuzzy expert system for heart disease diagnosis," in *Proceedings of international multi conference of engineers and computer scientists*, Hong Kong, 2010.
- [39] S. Kumar and G. Kaur, "Detection of heart diseases using fuzzy logic," *Int. J. Eng. Trends Technol. (IJETT)*, vol. 4, no. 6, pp. 2694--2699, 2013.
- [40] T. Kasbe and R. S. Pippal, "Design of heart disease diagnosis system using fuzzy logic," in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017.
- [41] N. Allahverdi, S. Torun and I. Saritas, "Design of a fuzzy expert system for determination of coronary heart disease risk," in *Proceedings of the 2007 international conference on Computer systems and technologies*, 2007.
- [42] K. K. Oad, X. DeZhi and P. K. Butt, "A fuzzy rule based approach to predict risk level of heart disease," *Global Journal of Computer Science and Technology*, vol. 14, no. 3, 2014.
- [43] S. Subbulakshmi, G. Marimuthu and M. N. Neelavathy, "A Fuzzy Logic Decision Support System for The Diagnosis of Heart Disease," *IOSR Journal of Engineering (IOSRJEN)*, vol. 8, no. 8, pp. 70-77, 2018.
- [44] N. Mishra and P. Jha, "A review on the applications of fuzzy expert system for disease diagnosis," *International Journal of Advanced Research in Engineering and Applied Sciences*, vol. 3, no. 12, pp. 28--43, 2014.
- [45] K. Ali, A. Keleş and U. Yavuz, "Expert system based on neuro-fuzzy rules for diagnosis breast cancer," *Expert systems with applications*, vol. 38, no. 5, pp. 5719--5726, 2011.
- [46] V. Balanica, I. Dumitrache, M. Caramihai, W. Rae and C. Herbst, "Evaluation of breast cancer risk by using fuzzy logic," *University Politehnica of Bucharest Scientific Bulletin, Series C*, vol. 73, no. 1, pp. 53-64, 2011.
- [47] K. Latha, B. Madhu, S. Ayesha, R. Ramya, R. Sridhar and S. Balasubramanian, "Visualization of risk in breast cancer using fuzzy logic in matlab environment," *International Journal of Computational Intelligence Techniques*, vol. 4, no. 1, p. 114, 2013.
- [48] K. Lavanya, M. S. Durai and N. Iyengar, "Fuzzy rule based inference system for detection and diagnosis of lung cancer," *International Journal of Latest Trends in Computing*, vol. 2, no. 1, pp. 165--171, 2011.
- [49] I. Saritas, N. Allahverdi and I. U. Sert, "A fuzzy approach for determination of prostate cancer," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 1, no. 1, pp. 1-7, 2013.
- [50] S. Kolhe, R. Kamal, H. S. Saini and G. Gupta, "A web-based intelligent disease-diagnosis system using a new fuzzy-logic based approach for

- drawing the inferences in crops," *Computers and Electronics in Agriculture*, vol. 76, no. 1, pp. 16--27, 2011.
- [51] A. Patel, J. Choubey, S. K. Gupta, M. Verma, R. Prasad and Q. Rahman, "Decision support system for the diagnosis of asthma severity using fuzzy logic," in *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS'12)*, 2012.
- [52] S. K. Anand, R. Kalpana, S. Vijayalakshmi, S. Hartley and B. Boucho-meunier, "Design and implementation of a fuzzy expert system for detecting and estimating the level of asthma and chronic obstructive pulmonary disease," *World Applied Sciences Journal*, vol. 23, no. 2, pp. 213--223, 2013.
- [53] N. Mishra, D. Singh, M. K. Bandil and P. Sharma, "Decision support system for asthma (DSSA)," *International Journal of Information and Computation Technology*, vol. 3, pp. 549--554, 2013.
- [54] K. Sundararaman, V. Seerkalan and K. Rangarajan, "Risk Level of Asthma and Chronic Obstructive Pulmonary Disease Through Design of an Intelligent Type-2 Fuzzy Expert System," *Instrumentation Measure Métrologie*, vol. 18, no. 6, pp. 583-590, 2019.
- [55] S. R. Damirchi-Darasi, M. F. Zarandi, I. Turksen and M. Izadi, "Type-2 fuzzy rule-based expert system for diagnosis of spinal cord disorders," *Scientia Iranica. Transaction E, Industrial Engineering*, vol. 26, no. 1, pp. 455--471, 2019.
- [56] M. A. Kadhim, M. A. Alam and H. Kaur, "Design and implementation of fuzzy expert system for back pain diagnosis," *International Journal of Innovative Technology & Creative Engineering*, vol. 1, no. 9, pp. 16-22, 2011.
- [57] H. Zarei, A. V. Kamyad and A. A. Heydari, "Fuzzy modeling and control of HIV infection," *Computational and mathematical methods in medicine*, vol. 2012, 2012.
- [58] A. Imianvan, U. Anosike and J. Obi, "An Expert System for the Intelligent Diagnosis of HIV/AIDs Using Fuzzy Cluster Means Algorithm," *Global Journal of Computer Science and Technology*, 2011.
- [59] T. Faisal, M. N. Taib and F. Ibrahim, "Adaptive Neuro-Fuzzy Inference System for diagnosis risk in dengue patients," *Expert Systems with Applications*, vol. 39, no. 4, pp. 4483--4495, 2012.
- [60] D. Saikia and J. C. Dutta, "Early diagnosis of dengue disease using fuzzy inference system," in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, 2016.
- [61] M. Alrashoud, "Hierarchical Fuzzy Inference System for Diagnosing Dengue Disease," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2019.
- [62] "What Is The Normal Body Temperature: Babies, Kids, Adults, And More," *Healthline*, 2020. [Online]. Available: <https://www.healthline.com/health/what-is-normal-body-temperature#temperature-range>.
- [63] "Measuring Your Peak Flow Rate," *American Lung Association*, 2020. [Online]. Available: <https://www.lung.org/lung-health-diseases/lung-disease-lookup/asthma/living-with-asthma/managing-asthma/measuring-your-peak-flow-rate>. [Accessed 1 March 2021].

An Optimization Approach for Multiple Sequence Alignment using Divide-Conquer and Genetic Algorithm

Arunima Mishra¹

Computer Science and Engineering
AKTU Lucknow, India

Sudhir Singh Soam²

Computer Science and Engineering
IET Lucknow, India

Bipin Kumar Tripathi³

Computer Science and Engineering
REC Bijnor, India

Abstract—Multiple Sequence Alignment (MSA) is a very effective tool in bioinformatics. It is used for the prediction of the structure and function of the protein, locating DNA regulatory elements like binding sites, and evolutionary analysis. This research paper proposed an optimization method for the improvement of multiple sequence alignment generated through progressive alignment. This optimization method consists of a fusion of two problem-solving techniques, divide-conquer and genetic algorithms in which the initial population of MSAs was generated through progressive alignment. Each multiple alignment was then divided vertically into four parts, three genetic operators were applied on each part of the MSA, recombination was done to reconstruct the full MSA. To estimate the performance of the method the results generated through the method are compared with well-known existing MSA methods named Clustal Ω , MUSCLE, PRANK, and Clustal W. Experimental results showed an 11-26% increase in sum_of_pair score (SP score) in the proposed method in comparison to the above-mentioned methods. SP score is the cumulative score of all possible pairs of alignment within the MSA.

Keywords—Multiple sequence alignment; divide; and conquer; genetic algorithm; optimization method

I. INTRODUCTION

Sequence alignment (SA) is the most common and essential task of bioinformatics. Pairwise SA is an alignment of two biological sequences where the similarity between the two sequences has been revealed through the alignment, few examples of SA are EMBOSS [1], BLAST[2], PSI-BLAST[3], and AlignMe[4], in the case when three and more sequences are aligned, is referred as multiple sequence alignment (MSA). The objective of the MSA is to arrange the sequences in a way that exposes the evolutionary connection between the biological sequences. The key applications of MSA are the identification of a protein family-like phylogenetic analysis and finding DNA regulatory elements. MSA is a well-known problem of combinatorics and its complexity is quite high, hence to get an exact solution is not practically possible for a large number of sequences [5], that is why most of the multiple alignment methods are heuristic and provide approximate solutions. Progressive alignment and iterative alignment are the two most applied approaches for MSA. The progressive alignment method is primarily based on the PSA in which pairwise alignment is done for all the possible pairs of sequences, a distance matrix is made that shows the dissimilarities between the sequences. A guide tree

is constructed through the distance matrix by any clustering algorithm. The guide tree displays the order of sequences to be aligned, most similar sequences are aligned first followed by the sequences of less similarity. Feng and Doolittle were the first who proposed a progressive alignment algorithm for MSA [6]. Many MSA methods based on the progressive alignment have been developed like CLUSTALW [7], MULTALIGN [8], CLUSTAL X[9]. The major disadvantage of this method is that the resulting MSA gets affected by the initial alignments so the position and length of gaps of aligned sequences can not be changed at a later stage.

Iterative alignment provides a solution to this problem through iteratively modifying the previously aligned sequences while keeps on adding the new sequences, few examples of iterative alignment are MAFFT [10,11], MUSCLE [12,13], and PRRP [14].

The machine learning area has been explored and several methods are applied to deal with the MSA problem. Ant colony optimization was applied by Chen et al [15] it was a partitioning approach that consists of three phases. Ant colony optimization was applied on each part and at last, all the parts were reassembled to get the solution.

Particle swarm optimization was combined with the Hidden Markov model to get the MSA by Rasmussen et al [16], they displayed improved results for protein sequences than the other HMM method for MSA like simulated annealing [17].

Reinforcement learning (RL) algorithms are used in solving the problem. Mircea et al [18] applied it the first time. The Q learning algorithm was applied along with the action-selection approach like softmax and epsilon-greedy for balancing the explore-exploit strategy. This strategy explores the solution space that may not provide instant high scores but may lead to a higher gain in the longer term. Exploitation is the application of information already gained by prior experiences. A good balance of exploration and exploitation helps to reach the optimum result in lesser time. Reza Jafri et al [19] used the deep Q learning method along with the actor-critic algorithm and experience replay method. They showed that their method has a speedy convergence. RLALIGN [20] is a pure RL-based algorithm for MSA.

Genetic algorithm (GA) is a type of iterative method, analogous to the theory of natural evolution. It generates many solutions at each stage every solution is attached with a fitness function that describes the goodness of that solution. The genetic operators like mutation operators and recombination operators are applied to the selected entities at each stage iteratively until the result converged to the best possible fitness score. Due to the MSA's discrete nature, GA is well suited to this problem. SAGA [21] is a famous MSA method developed by Higgins and Notredame, based on the GA. It attempts to get the MSA by the number of complex genetic operators. One more approach was proposed by Nazneen et al [22] in which the initial population was generated through randomly produced subtrees and then by shuffling of those subtrees. MSA-GA [23] is another method in which the initial population was produced through dynamic programming and then Genetic operators were applied to it to get the next generation population iteratively.

This paper suggests an approach named Genetic algorithm-based optimization with divide and conquer method (GAODC) which is a fusion approach of two problem-solving techniques namely a genetic algorithm and divide and conquer methodology. In this approach, the Sum_of_Pair score is used as a fitness function. It divides the MSAs into four parts and two operators namely insertion mutation and deletion mutation are applied to those parts. The fitness score is calculated for each part and recombination is done between the parts of two MSAs starting with the MSAs having the highest fitness score.

The most distinguishing feature of this methodology is that the recombination of MSAs is based on the fitness score of the individual parts of MSAs. Recombination between the MSAs with high fitness scores has a great chance of construction of new MSA with higher fitness scores. To evaluate the performance of GAODC, it is compared with other popular MSA methods, namely PRANK[24], CLUSTAL Ω , MAFFT, and MUSCLE. BAliBASE 3.0 [25] database is used for the evaluation of the method. Sum_of_pair score and total_column score are the two objective functions that are used as an evaluation criterion for all the MSA methods.

The rest of the paper is principally divided into eight main sections: Section II presents the basic definitions of the progressive alignment, Divide-conquer approach, and Genetic algorithm. The methodology of GAODC is explained in Section III, Section IV mentioned the fitness function and the scoring scheme. Datasets used in the method are explicated in Section V. Results generated through GAODC and other existing methods are compared in Section VI, Section VII summarizes the conclusions, and Section VIII lists out the references cited in the paper.

II. PRELIMINARIES

This section contains the preliminaries that are used in the proposed method. Section A explains the progressive alignment technique that is applied to generate the initial population, section B and C contains the basic steps of divide and conquer and genetic algorithm respectively.

A. Progressive Alignment

Progressive alignment is a basic technique of multiple sequence alignment it starts with the pairwise sequence alignment of any two sequences and then the third sequence is aligned, and this process continues till all the sequences get aligned. This method does not guarantee optimal alignment, but it is a very fast method for MSA. Following are the main steps of progressive alignment:

1) Make the distance matrix for $M(M-1)/2$ pairs of sequences of M sequences.

2) Make a guide tree with the help of the matrix using clustering algorithms like neighbor-joining [26] and UPGMA[27], which shows the order of sequences to be aligned.

3) Add the sequences into the alignment starting from the sequences added first followed by other sequences added to the guide tree. An example of a guide tree is shown in Fig. 1 for the sequences S1, S2, S3, and S4. As depicted in the figure, S1 and S2 will be aligned first followed by S3 and S4.

B. Divide and Conquer

Divide and conquer is a recursive problem-solving approach. It breaks the complex problem into smaller subproblems of similar type till the subproblems are converted to simple problems that can easily be solved then each subproblem is solved and combined to obtain a complete solution. Three main steps of the divide and conquer method are divide: in which the problem is divided into subproblems of the same kind, the second step is to conquer, that solves the subproblems recursively and the final step is to combine, where all the solutions are combined to achieve the final solution of the entire problem.

C. Genetic Algorithm

Genetic algorithms (GA) are analogous to the process of evolution where the best individuals are chosen to produce next-generation offspring. The main steps of GA are as follows:

- 1) Generation of the initial population.
- 2) Calculate the fitness function for each entity within the population.
- 3) Choose some individual as parents.
- 4) Apply genetic operators on them.
- 5) Produce the next generation with the help of some recombination of the previous generation.
- 6) Repeat steps 2 to 5 until the stop criterion.
- 7) End.

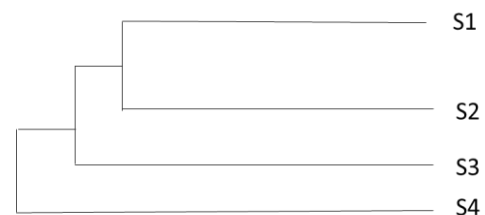


Fig. 1. The Guide Tree Displays the Order of the Sequences to be Aligned in Progressive Alignment.

III. METHOD

Genetic algorithm-based optimization with divide and conquer (GAODC) consists of the following main steps: population initialization, division, mutation, and recombination.

After generating the initial population with the help of progressive alignment and random insertion of gaps, the division of each individual into four upright parts is performed, Fig. 3 shows the vertical division of the MSA. Mutation operators are applied on each part of the MSA and two types of recombination (One Point and Two Point recombination) are performed to generate a next-generation population. Mutation operations are done on each part of the MSA instead of the whole MSA and one point and two-point recombination are achieved based on the fitness score. Application of mutation operators on the vertical parts of MSA instead of full MSA and the fitness score-based recombination are the main features of the method. This vertical division is done after the initial population generated, two mutation operators, namely, insertion mutation operator and deletion mutation operators are applied on each segment of the MSA to achieve a better alignment score.

A. Gap Insertion Mutation

a gap insertion mutation operator is introduced that picks an MSA from the population and creates a gap randomly at each row, the changes are retained if fitness improves. Fig. 2 illustrates the example of gap insertion mutation for the following four sequences:

S1: ABV, S2: BV, S3: AV, and S4:ABV

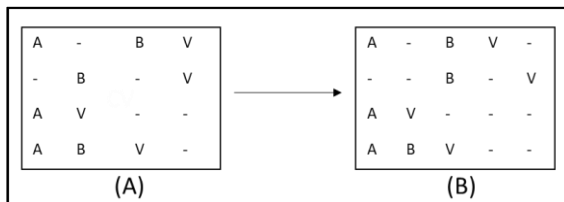


Fig. 2. Depicts Insertion Mutation Applied on Alignment A and the Resulting Alignment B.

B. Gap Removal Mutation

A gap elimination mutation operator is designed that picks an MSA and selects the positions randomly to eliminate the gap. If the selected positions are not a gap then move forward till a gap is found and delete the gap. If no gap is found, then go back and delete the first gap found. Retain the changes if the fitness score is increased. Fig. 3 illustrates the example of deletion mutation operation on MSA (A) for the sequences S1, S2, S3, and S4.

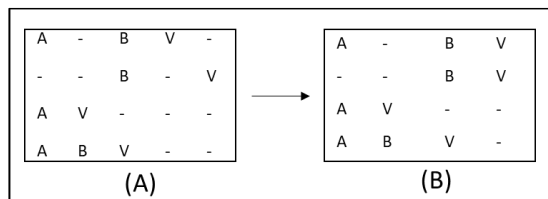


Fig. 3. Depicts Deletion Mutation Applied to Alignment A and the Resulting Alignment B.

The key feature of the algorithm is the division step which takes place before the mutation operators are applied and the mutation is done on the part of the MSAs instead of the full MSA. The crux of this algorithm is that this strategy increases the randomness and the recombination between the parts of the parents based on the fitness score, enhances the chance of getting a new MSA with a higher score. Following are the main steps of the algorithm-

1) *Population initialization*: The initial population is produced with the help of the progressive alignment technique which consists of the following steps:

A. Compute the distance between each possible pair of sequences using the formula:

$$d(Q_i, Q_j) = 1 - \{ M(Q_i, Q_j) / \min(l_i, l_j) \} \quad (1)$$

Here $M(Q_i, Q_j)$ - Number of matches between the i th sequence and j th sequence.

l_i and l_j – Sequence length of the i th sequence and j th sequence, respectively:

B. A guide tree is constructed to get the order of sequences to be aligned.

C. Sequences are aligned in the order directed by the guide tree. Most identical sequences are aligned first followed by the distant sequences. Three types of alignment are possible here

- a. Sequence to sequence alignment
- b. Group of aligned sequences to a sequence
- c. Group to group.

After generating the MSA through progressive alignment, random gap insertions are done to generate the initial population of size n .

2) *To generate the next generation alignments these steps are used*: Division: Divide the individual MSA into four upright parts for example, from MSAs 1 to n , it will be - a_1, b_1, c_1, d_1 to a_n, b_n, c_n, d_n . Deletion mutation operators and insertion mutation operators are applied on each part of the MSA individually and changes are saved with the highest fitness score of each part. Now we have four parts of each parent, fitness score of $(a+b)$, and $(a+b+c)$ parts are calculated for all the MSAs. Fig. 4 illustrates the example of the division process for four sequences S1: ABVKWSPNVS, S2: BVKWSNS, S3: AVKSPV, and S4: ABVKSYS. Now two types of recombination are done to produce the next generation alignments, one-point recombination, and two-point recombination. An illustration of one-point recombination for the above example is shown in Fig. 6, 7 and 8 whereas Fig. 5 depicts the two-point recombination for the above-mentioned example.

One-point recombination: It contains the following steps:

1) (a) part of one parent with $MaxScore(a)$ will be combined with the $(b+c+d)$ part of the other parent with $MaxScore(b+c+d)$.

2) (a+b) part of one parent with MaxScore (a+b) will be combined with the (c+d) part of the other parent with MaxScore(c+d).

3) (a+b+c) part of one parent with MaxScore(a+b+c) will be combined with the d part of another parent with MaxScore(d).

4) Continue Steps 1, 2 and 3 with other parents with the next maximum scores.

5) Evaluate the fitness function for all new MSAs.

Two-point recombination: Two-point recombination contains the following steps:

1) Part a and c of one parent having MaxScore(a+c) are recombined with the b and d part of the other parent with MaxScore(b+d).

2) Continue step 1 with other parents with the next maximum scores.

3) Evaluate the fitness score for all new MSAs.

Elitism: This is a popular approach of the genetic algorithm where the individual with the highest fitness value of that generation, is passed as it is to the next generation so that we may not lose the best solution at any stage.

New generation: The creation of a new generation is formed with the selection of the best distinct half of the collective parents, and children who are generated through the mutation and crossovers. The key feature of the method is that selection of parents is not random but it is based on the fitness score of the part to be recombined with the part of the other parent and this increases the possibility of getting a higher score after recombination and safeguard a good balance between exploitation and exploration. The new generation formation (shown in Fig. 9) is considered as the parent population of the next generation and therefore the process of evolution continues.

Termination condition: The best score and its corresponding MSA have recorded in each generation if there is no improvement in the solution till 100 MSA then the execution of the algorithm will be ended.

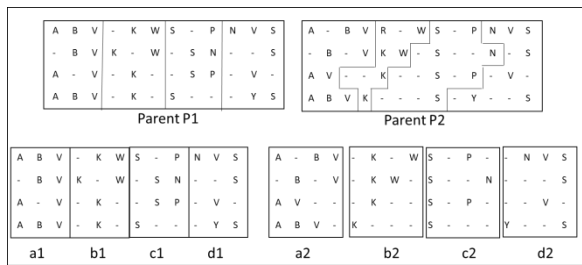


Fig. 4. Depicts the Division Process of MSAs in Four Parts Namely a,b, c, and d.

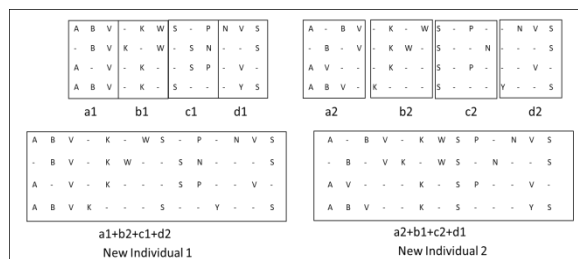


Fig. 5. Two-point Crossover after Applying Insertion and Deletion Mutation on a, b, c, and d Parts of the MSAs.

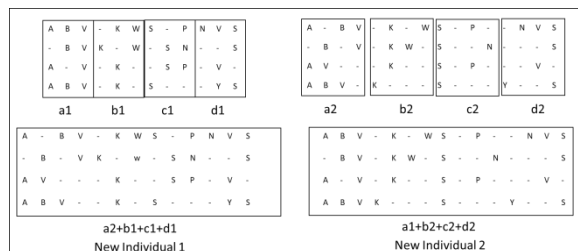


Fig. 6. One-point Crossover on (a) Part of One Parent with MaxScore (a) Combined with the (b+c+d) Part of the other Parent with MaxScore (b+c+d), after Applying Insertion and Deletion Mutation on a, b, c, and d Parts of the MSAs.

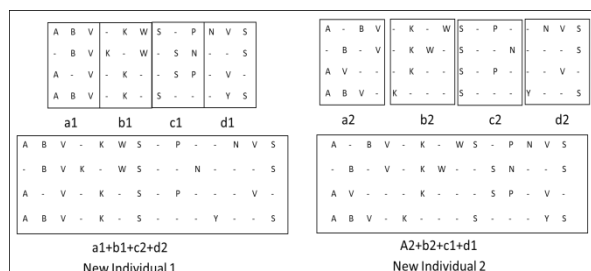


Fig. 7. One-point Crossover on (a+b) Part of One Parent with MaxScore (a+b) Combined with the (c+d) Part of the other Parent with MaxScore (c+d), after Applying Insertion and Deletion Mutation on a, b, c, and d Parts of the MSAs.

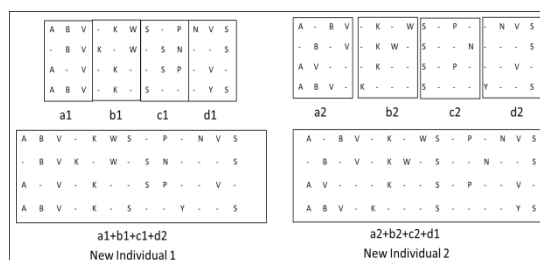


Fig. 8. One-point Crossover on (a+b+c) Part of One Parent with MaxScore(a+b+c) Combined with the (d) Part of the other Parent with MaxScore (d), after Applying Insertion and Deletion Mutation on a, b, c, and d Parts of the MSAs.

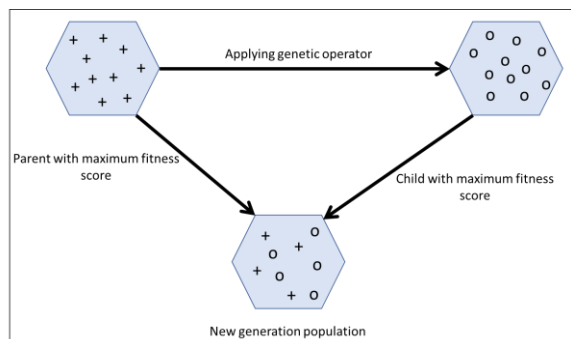


Fig. 9. Depiction of Construction of New Generation Population.

IV. FITNESS FUNCTION AND SCORING SCHEME

The fitness function in a genetic algorithm comprises all the parameters of the particular problem and estimates the solution that how much it is close to the actual solution. In the MSA problem, the most reliable scoring scheme Sum_of_Pair score (SP score) is taken as the fitness function, it is represented by the equation 2:

$$S = \sum_{j=1}^{n-1} \sum_{k=j+1}^n S(j, k) + \text{Gap-Penalty} \quad (2)$$

Here $S(j,k)$, is the sum of the pair score of j th and k th sequences, and 'n' is the total number of sequences. The sum_of_pair score for all $n(n-1)/2$ pairs of the biological sequences is computed and added. For the match/mismatch score the BLOSUM 62 matrix is used.

The gap penalty is a fine incurred for inserting a gap in the process of MSA. The affine gap penalty is applied to compute the fine, represented by equation 3:

$$Gp = A + B (t-1) \quad (3)$$

Where A is the Gap Opening penalty, B is the gap extension penalty and t is the number of consecutive gaps in a row.

To calculate the quality of MSA methods one more parameter total_column_score (TC score) is being used. TC score evaluates the ability of the MSA methods to align all the residues appropriately in each column. Mathematically it is defined in equation 4:

$$S = \sum_{i=1}^d \begin{cases} 1 & \text{if } T_i = R_i \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

Here d is the length of the MSA, T_i and R_i are the i th columns of test MSA and reference MSA respectively if the column of test MSA matched completely it will return '1' else it will return '0'. The summation of all the values for each column divided by the number of total columns gives the value of TCscore as shown in equation 5.

$$\text{TCscore} = S/d \quad (5)$$

V. DATABASE

To evaluate the performance of the proposed method the dataset BALiBASE V3.0 (<http://www.lbgi.fr/balibase/>) was chosen. It is a commonly used benchmark database of protein sequences. It contains the set of protein sequences and their corresponding reference MSAs. It comprises an application

BALiscore that calculates the SPscore and TCscore of test MSAs with the comparison of reference MSAs, its scores vary in the range of 0 to 1. If the test MSA is identical to the reference MSA the value of BALiscore is '1' and if the test MSA not at all matches the reference MSA BALiscore is '0'.

BALiBASE v3 contains six different groups of protein sequences namely RV11, RV12, RV20, RV30, RV40, and RV50. RV11 has 38 sets of very divergent protein sequences that are equidistant and have <20% identity. RV12 was constructed by 44 sets of sequences of 20%- 40% similarity. RV20 formed with 41 sets of sequences having few orphan sequences while all other sequences are having <40% sequences. RV 30 consists of 30 sets of sequences from different families having <25% of identity among the families. RV 40 is a set of 49 sets of protein sequences with a large number of insertions whereas RV 50 is formed by 16 sets of sequences having a large number of internal insertions with < 20% identity.

VI. RESULTS AND DISCUSSION

To evaluate the results of GAODC, its results for benchmark database BALiBASE 3.0 are compared with four other MSA methods namely Clustal Ω , MUSCLE, Clustal W, and PRANK. Parameter values for GOADC are given in Table I. Two criteria SPscore and TCscore are used to evaluate the quality of the MSAs. The results are shown in Table II.

Fig. 10 depicts the sp scores of five methods across the six datasets for which the analysis is conducted. Fig. 11 illustrates the TC Score across the same datasets used to calculate SP Score for all five methods.

TABLE I. PARAMETERS USED IN GAODC

Name	Value
Population Size	100
Substitution Matrix	BLOSUM 62
Gap Penalty	Gap opening -3, Gap extension-2

TABLE II. RESULTS OF SP SCORE AND TC SCORE OF EACH METHOD ACROSS ALL SIX DATASETS

Methods		GAODC	CLUSTAL Ω	MUSCLE	CLUSTAL W	PRANK
RV11	SP	0.548	0.452	0.442	0.5	0.354
	TC	0.25	0.247	0.228	0.229	0.168
RV12	SP	0.878	0.826	0.827	0.865	0.737
	TC	0.763	0.683	0.679	0.717	0.548
RV20	SP	0.855	0.772	0.766	0.852	0.7
	TC	0.15	0.31	0.25	0.22	0.17
RV30	SP	0.82	0.68	0.65	0.73	0.51
	TC	0.33	0.37	0.25	0.28	0.18
RV40	SP	0.845	0.756	0.726	0.789	0.6
	TC	0.402	0.428	0.338	0.398	0.244
RV50	SP	0.747	0.681	0.724	0.742	0.55
	TC	0.481	0.417	0.344	0.312	0.245

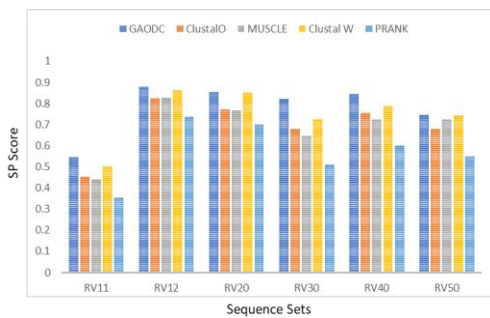


Fig. 10. Comparative Results of SPscore of each Method Across all Six Datasets.

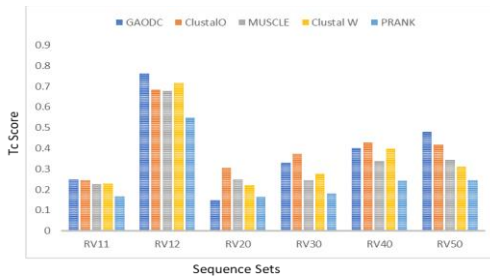


Fig. 11. Comparative Results of TCscore of each Method Across all Six Datasets.

Results show that the overall performance of the method is better than other methods. SP score of GAODC is approximately 13% higher than all other methods across six databases. The TC score of the GAODC method is highest among all other methods for four datasets including RV11, RV12, RV40, and RV50. In the case of datasets RV 20 and RV 30, which contain orphan sequences and groups of sequences having different families respectively, the TCscore of GAODC is marginally less by approximately 1.7% than the other methods.

High SP Score and TC Score suggest that GAODC generate better quality MSAs as compared to other methods used in this analysis.

VII. CONCLUSION

This paper proposes a method for the MSA of biological sequences that is a combination of two problem-solving techniques divide-conquer and genetic algorithm. As part of this method, the recombination method is applied where the MSAs are recombined based on the SP score of the parts of each MSA thus increasing the possibility of getting the most optimum MSA. Results show that our method outperformed the other widely used MSA techniques on SPscore criteria.

REFERENCES

- [1] Rice P, Longden I, Bleasby A, "EMBOSS: the European molecular biology open software suite", Trends Genet. 2000; 16:276–7.
- [2] Johnson M, Zaretskaya I, Raytselis Y, Merezuk Y, McGinnis S, Madden TL, "NCBI BLAST: a better web interface", Nucleic Acids Res. 2008;36: W5–9.
- [3] Altschul SF1, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
- [4] Marcus Stamm, René Staritzbichler, Kamil Khafizov, and Lucy R. Forrest, "AlignMe—a membrane protein sequence alignment web

- server", Nucleic Acids Res. 2014 Jul 1; 42(Web Server issue): W246–W251.
- [5] F. Sievers, A. Wilm, D. Dineen, et al., "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", Molecular Systems Biology, vol. 7, article 539, 2011.
- [6] Feng DF, Doolittle RF, "Progressive sequence alignment as a prerequisite to correct phylogenetic tree", 1987, J Mol Evol. 25 (4): 351–360.
- [7] Thompson JD, Higgins DG, Gibson, ". CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice", Nucleic Acids Res. 1994;22:4673–80.
- [8] Corpet F., "Multiple sequence alignment with hierarchical clustering". Nucl Acids Res. 1988.
- [9] Francois Jeanmougin et al "Multiple sequence alignment with Clustal X " Trends in biochemical sciences, 1998.
- [10] Katoh K, Standley DM, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability", Mol Biol Evol. 2013; 30:772–80.
- [11] Katoh K, Misawa K, Kuma K, Miyata T, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", Nucleic Acids Res., 2002;30:3059–66.
- [12] Edgar RC, "MUSCLE: multiple sequence alignment with high accuracy and high throughput", Nucleic Acids Res. 2004;32:1792–7.
- [13] Edgar RC, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity", BMC Bioinformatics. 2004; 5:113.
- [14] Gotoh O, "Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments", J Mol Biol 1996, 264: 823–838.
- [15] Chen Y, Pan Y, Chen L, Chen J (2006) Partitioned optimization algorithms for multiple sequence alignment. In: Proceedings of the 20th international conference on advanced information networking and applications, pp 618–622.
- [16] Rasmussen TK, Krink T (2003) Improved hidden Markov model training for multiple sequence alignment by a particle swarm optimization-evolutionary algorithm hybrid. BioSystems 72:5–17.
- [17] Jin Kim, Sakti Pramanik, Moon Jung Chung, "Multiple sequence alignment using simulated annealing", Bioinformatics, Volume 10, Issue 4, July 1994.
- [18] M.I. Bocicor, I.G. Mircea, and G. Czibula. "A novel reinforcement learning-based approach to multiple sequence alignment, Information Sciences, 2014.
- [19] Reza Jafari, Mohammad Masoud Javidi · Marjan Kuchaki Rafsanjani, "Using deep reinforcement learning approach for solving the multiple sequence alignment problem", 2019, Springer Nature Switzerland.
- [20] RK Ramakrishnan, J.Singh and M. Blanchette, "RLALIGN: A Reinforcement Learning Approach for Multiple Sequence Alignment", IEEE 18th International Conference on Bioinformatics and Bioengineering, 2018.
- [21] Notredame C, Higgins DG (1996) SAGA: sequence alignment by genetic algorithm. Nucl Acids Res 24:1515–1524.
- [22] Naznin F, Sarker R, Essam D (2011) Vertical decomposition with genetic algorithm for multiple sequence alignment. BMC Bioinformatics 12:353.
- [23] Gondro C, Kinghorn BP (2007) A simple genetic algorithm for multiple sequence alignment. Genet Mol Res 6:964–982.
- [24] Ari Loytynoja, "Phylogeny-aware alignment with PRANK", multiple Sequence Alignment Methods pp 155-170.
- [25] Thompson JD, Koehl P, Ripp R, Poch O, "BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark", Proteins. 2005 ;61(1):127-36.
- [26] Saitu N, Nei M, "The neighbor-joining method: a new method for reconstructing phylogenetic trees", Mol Biol Evol. 1987 Jul;4(4):406-25.
- [27] I.Gronau and S.Moran, "Optimal implementations of UPGMA and other common clustering algorithms", Information Processing Letters, vol.104, no.6, pp.205–210,2007.

Proposed Design of White Sugar Industrial Supply Chain System based on Blockchain Technology

Ratna Ekawati¹

Agro-industrial Engineering Graduate Students IPB
Staff Industrial Engineering Untirta
Cilegon, Banten, Indonesian

Yandra Arkeman², Suprihatin³, Titi Candra Sunarti⁴

Agro-Industrial Engineering, IPB University
Bogor, West Java
Indonesian

Abstract—The white crystal sugar agro-industry is an industry with dynamic characteristics characterized by a sustainable relationship between actors ranging from farmers to consumers. An inefficient supply chain system will affect the flow of products, information, and finance because many actors are involved and have influence. Hence, complicating the system in the tracking process flow, product flow and creating problems that occur in business processes. The main objective of this research is to propose the design of an integrated white crystal sugar agro-industrial supply chain system based on blockchain technology so that it can increase competitiveness in realizing food security and resilience; by proposing a search for the problem of mismatches that occur along the supply chain from upstream to downstream. The variables that will be identified in the supply chain flow include quality, quantity, and price, with the suitability of transaction information data ranging from farmers, sugar factories, warehouses, distribution, retailers to the final consumer. It is hoped that consumers will feel happy to consume trusted local sugar with the best safety and quality, as well as ensure transparency of information between actors. Previous traditional methods, which were still centralized, would be transformed into decentralized information, to create trust among stakeholders. With a blockchain-based traceability architecture design, it is hoped that the proposed design can be implemented in the white crystal sugar agro-industry.

Keywords—Blockchain technology; supply chain; white crystal sugar

I. INTRODUCTION

White crystal sugar is one of the strategic commodities in the agricultural industry [1], therefore it is necessary to maintain both quality and quantity to achieve national food security and security. Quality improvement based on Icumsa 81-200 IU (SNI 3140.3: 2010), because the sugar produced during the milling period is not higher than 200 IU, causing waste in the production process.

The white crystal sugar agro-industry is an industry with dynamic characteristics characterized by a sustainable relationship between actors ranging from farmers to consumers. The low average efficiency of sugar mills based on plantation performance is 61.78% and factory performance 63.05% [2], the performance of sugar mills in Yogyakarta province is 72.338% [3], the performance of partner farmers is 75.86% [4]. Raw sugar production tended to decline over the last three years, in 2017 the production of raw sugar was only 2.21 million tonnes.

Sugar supply chain problems are very complex [5], [6], multi-stakeholder, long, high uncertainty, lack of coordination and integration, especially on the flow of data information distribution from farmers, factories. up to the consumer table [7], [8], [9]. This causes the end consumer not to have accurate and precise information regarding the quality and quantity of sugar. Because consumers only know that sugar is sweet, but do not know the meaning contained in the quality of the sugar, the difference in the color of sugar, and the shape of the grains on the market.

Inefficiency does not only occur in the main stakeholders but also occurs for supporting stakeholders such as associations that play a role and state institutions that play a role in monitoring the smooth running of production in terms of quantity and quality. So that sometimes power or wealth in terms of capital makes stakeholders strong who control data and game information in the sugar supply chain.

Less integrated product, financial and information flows can also affect supply chain performance [5]. Without coordination, it will be difficult to synergize data from upstream to downstream of the supply chain [10]. Therefore, an optimal, efficient and transparent supply chain system design is needed to improve the performance of business processes in the sugar agro-industry [9],[11], [12].

In the current era, supply chain problems also require traceability to be able to identify products, processes, and environmental characteristics, decision-making information, and overall system analysis [13], [14]. The development of the agro-industrial supply chain until the 5.0 era [15] was market-oriented, which had consumer feedback, and transparency about the quality, quantity, and process.

The proposed application of innovative technology is blockchain technology which has the characteristics of decentralization, immutability, reliability, and transparency. Blockchain technology records all data transaction information in each supply chain that can help and represent traceability assets [16], [17]. Digital product information such as details of agricultural origin, batch numbers, factory data and processing, expiration date, storage temperature, and delivery details are digitally related to food items and information at each step of the process [18].

II. LITERATURE REVIEW

Changes in business systems by applying technology can bring strong trust and transparency among supply chain actors in increasing added value and maximizing economic benefits [19], [20]&[21]. In transactions between actors, so that the level of trust can be guaranteed, a new technology called blockchain is introduced, which is accompanied by a smart contract that contains rules or agreements for interacting. To avoid fraud and inaccuracy in payments, the distribution of data information between supply chain actors is created by creating an agreement between the actors who, if violated, will be given a penalty or sanction.

Agricultural products that have implemented blockchain-based on previous research are [22] proposed the application of blockchain to wine-based products with web-based RFID, [23] used a quantitative questionnaire for all actors involved in Burundi's coffee proposing blockchain technology based on the Technology Acceptance Model (TAM Hyperledger for Desktop Grain Controller (GDPA) Applications and GEBN Blockchain servers on grain products such as corn, soybeans, wheat in Brazilia apply digital blockchain contracts to all rice supply chain stakeholders in India, [24] applying the Ethereum blockchain and smart contracts for traceability of soybeans, [25] designed a blockchain business model that would help reduce logistics costs and optimize fresh vegetable supply chain operations, and proposed traceability to cocoa with digital watermarking [26].

Previous research is based on the study of Aung & Chang (2014) [14] that traceability is closely related to safety, health, and quality in the food supply chain industry [17], Tian (2016) develops supply chains. blockchain verification system with RFID which aims to identify [27], monitor, and check the journey of industrial products across all actors. Galvez & Mejuto (2018) blockchain is used to store data records with the role of sensors and IoT [28]. Blockchain according to Kamilaris et al. (2018) can build a reliable, transparent, and sustainable food supply chain environment, by integrating key stakeholders [29].

Following Fig. 1 below the research roadmap that will be carried out by researchers is based on previous reference papers regarding supply chain design, sugar supply chain, traceability, and blockchain technology. The researcher determined that the grand design of the study was a blockchain-based integrated white crystal sugar agro-industrial supply chain design technology that can increase competitiveness in realizing food security and resilience.

A search is carried out if there is a quality mismatch and an inaccurate amount of white sugar. Traceability of the sugar supply chain is carried out with stakeholders who play the main role based on the needs of each stakeholder from upstream to downstream. The technology that has been used and which will be used in the design of a supply chain system that aims to improve the performance of the sugar supply chain.

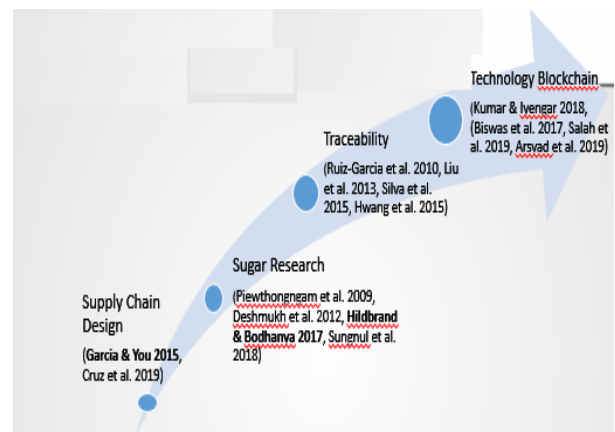


Fig. 1. Research Roadmap.

III. METHODOLOGY

The big objective of this research is to design an integrated white crystal sugar agro-industrial supply chain system based on blockchain technology that can increase competitiveness and realize food security and resilience. The steps in the flow of the research process are depicted in Figure 2. Several variables and data are used to produce research output. Data variables and supply chain systems from upstream to downstream are based on data in product flow, information flow to financial flow so that they become part of a decentralized blockchain consortium that uses smart contracts with the Ethereum platform.

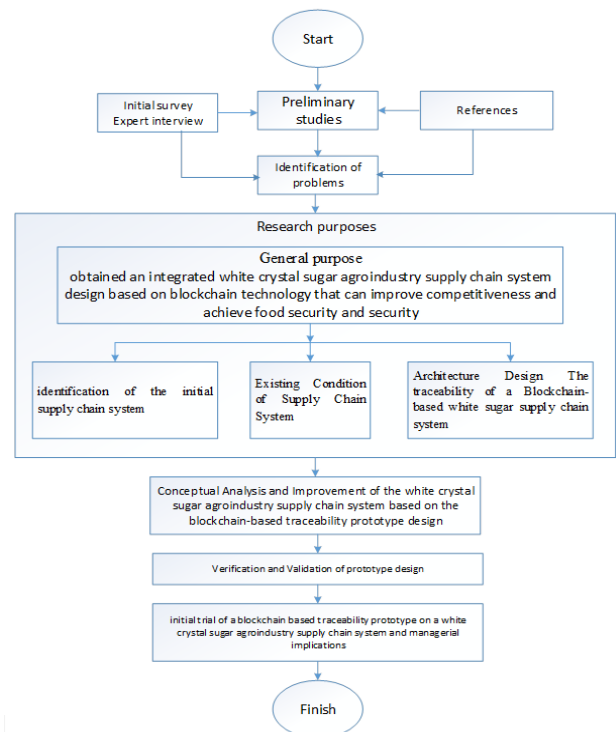


Fig. 2. Research Flow Chart.

Ethereum is a decentralized public blockchain network platform running smart contracts according to programming without any downtime or fraud. The blockchain work analogy that will be designed is that the actors participating in the supply chain system of the white crystal sugar agro-industry conduct or request transactions represented by each block, which will then be distributed to a peer network consisting of computers or nodes, where the nodes will validate the blocks and the transaction requested. If there are new transactions a new block will be created which will be added to the blockchain network and so on until the transaction is uploaded.

Specific objectives are stages to support the main objective in identifying the supply chain system of white crystal sugar agro-industry which consists of

- Identify the main factors influencing the flow of the supply chain and analyze their involvement such as farmers, factories, distributors, retailers, and consumers.
- Identify business processes and work activities that occur from upstream to downstream of the chain by collecting them based on literature reviews and also direct or indirect interviews based on preliminary surveys to sugarcane producing areas. The initial process of this identification will result in the supply chain system conditions of the existing white crystal sugar agro-industry.
- Identifying problem variables (critical) in the supply chain system of sugar agro-industry based on literature reviews, expert interviews (both from main actors and supporting actors regarding critical problems that often occur, such as quality, quantity, and price.

Traceability architecture design based on blockchain technology

- Based on the existing conditions that have been made, identification of business processes and activities that are more specifically focused on tracing the variables of critical problems that occur can be designed so that a scheme can be designed in the traceability architecture of the white crystal sugar agro-industrial supply chain system.
- Traceability architecture design is based on literature reviews (books, journals, reading articles related to supply chain and traceability) and conducting live discussions during the initial survey or through online discussions.
- Verification and validation of traceability architectural designs based on related statistical methods and including experts in decision making.
- Analyze the capabilities of the traceability scheme based on the requirements desired by an efficient supply chain system. Identification of integrated information flows data from upstream to downstream of the chain so that the data is timely and accurate.

Design a blockchain-based prototype based on traceability in the supply chain system with a description using UML. Based on the following steps:

- Designing and integrating transaction database information into the application architecture.
- User interface system design.
- Designing detailed blockchain-based prototypes.
- Verification and validation of blockchain-based prototypes based on traceability architecture.
- Initial trials before implementing the supply chain managerial system of white crystal sugar agro-industry.

Identification of the supply chain system configuration is based on the principles of the Supply Chain Operations Reference (SCOR) (Vorst 2014)[8], which is explained through four main elements, namely: a chain structure that describes the scope, roles of chain members and the agreements that make up the chain. A chain business process is a structured and measurable series of business activities to produce optimal output for consumers. Chain network management describes the coordination between actors in carrying out business processes in producing white crystal sugar products.

The condition of the existing system is based on a qualitative descriptive method that is supported by the opinions of sources, practitioners, initial field observations, and literature studies, and direct interviews through communication tools. The design of the supply chain system for white crystal sugar agro-industry will be modeled using the Unified Modeling Language (UML).

The prototype will be designed and implemented to facilitate transactions using the decentralized Ethereum Virtual Machine (EVM) system that allows tracking transactions between actors in the supply chain system, to build a reliable transparency process through the supply chain system from upstream to downstream to increase efficiency and productivity actors who play a role in the system as well as the quality of white crystal sugar products based on blockchain technology.

The system to be proposed is expected to provide transparency regarding the amount of production or availability of sugar cane, sugar, for each of the actors involved. Transparency in purchase prices among actors ranging from farmers to end consumers, and the quality of products in end consumers by improving their health and safety in consuming sugar according to the details of the composition of sugar contained in the packaging, coming from areas where sugarcane has been processed into white crystal sugar and also the date the expiry date. Whereas farmers can find out to what extent the sugarcane they plant and care for is consumed by consumers in which areas. So that the welfare of all actors involved in the supply chain system can be realized.

IV. ANALYSIS AND DISCUSSION

The globalization of trade causes supply chains to become more complex, even when the real relationship between

stakeholders is disharmony, and there is a lack of coordination, traceability of objects through the network is increasingly needed. Many sugar companies were forced to close due to fraud that occurred due to a lack of control over the production process, starting from the contract process with sugar cane farmers, the sugarcane weighing process, to the process of milling sugar cane into sugar.

However, there are also not a few sugar companies that still survive in producing sugar to remain stable and even get bigger due to financial support from the private sector. Delays and limitations in increasing inefficient supply chains, due to traditional flows and upstream to downstream data information transfers are still centralized. Inefficiencies that occur, such as the delivery of information that is often late and there is inequality of information received, dishonesty of data providers, and changing data arbitrarily so that companies cannot control the flow of information properly because the data is not real, not timely and reliable.

Traditional databases use a client-server network architecture. Here, the user (known as the client) can change the data, which is stored on a centralized server. Fixed database controls with defined authorities that authenticate client credentials before granting access to the database. This authority is in charge of database administration, if the security authorities are compromised, the data can be changed, or even deleted.

Fig. 3 explains the existing conditions that occur in the sugar agroindustry supply chain system. The main actors involved in the supply chain starting from farmers, operators of sugarcane yards (cutting and transporting), sugar factories, warehouses, corporate centers, distributors, retailers to consumers. While the general production process [30] is divided based on sections such as agronomy which contains data on contracts with farmers, land and agricultural classification data, charcoal data based on plantation area and area, sugarcane category and variety data, and data at the time of milling sugarcane per day.

In the hauling section, the available data are felling schedule data based on transport orders, transport data, data on

the number of incoming and weighed sugarcane. On the scale, there is data on the net weight of sugarcane (net weight - the empty weight of the truck) and data on tare weight (weight of sugarcane transported by crane). Data on the milling and QC processes consist of data on the classification of quality of sugarcane-based on (physical, cleanliness) and sugar classification data based on SNI.

The data warehouse that is centralized at the head office consists of data on the amount of sugar production, production time data, price data, batch data, and logistics operators, while distributors send delivery orders as well as retailers and consumers consisting of data on quantity, time, the accuracy of delivery and order recipients.

The problems that occur in the supply chain of white crystal sugar agro-industry currently require transparency and traceability for the development of a sustainable supply chain. To integrate problems into one major problem, such as the conflict between chain actors caused by distrust, data corruption, or data forgery that can harm all actors, product safety, product defects during storage and packaging, and other problems. Therefore, the researcher proposes a blockchain-based sugar agro-industrial supply chain system design architecture as shown in Fig. 4 which has the advantage of tracking problems that occur in the supply chain to be resolved quickly, transparently, resiliently, and fairly. Because the distribution of data that has been recorded in the block cannot be changed without the consent of the consensus involved.

All of these problems will force us to explore with the right approach to trace the origin, distribution, product status, expiry time, distribution time, and storage along the long and complex sugarcane supply chain. This is because tracking the existence of sugarcane commodities until sugar products are consumed by consumers through the chain, can improve the relationship between upstream farmers and downstream consumers, and it is also expected to control the quality of products produced by sugar factories.

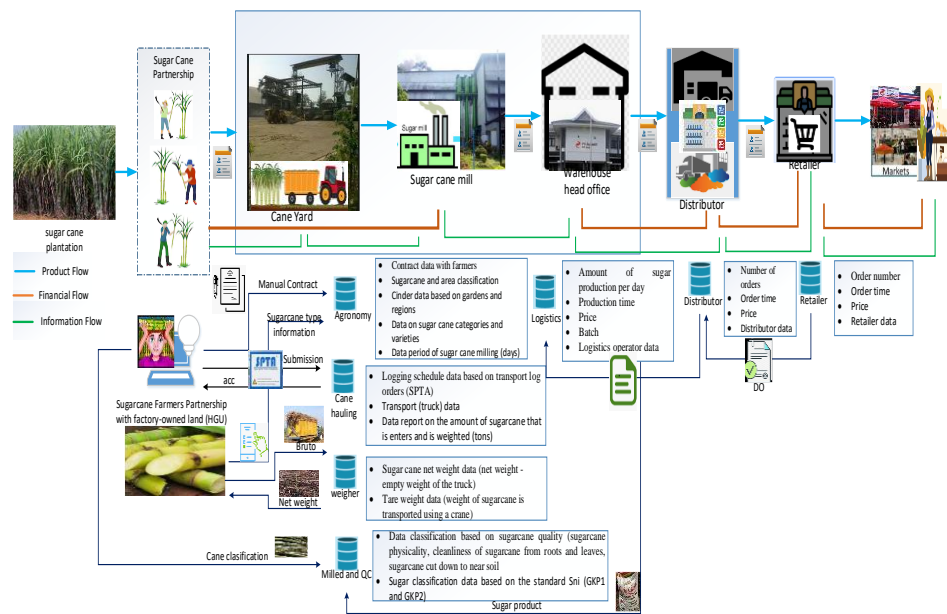


Fig. 3. The Existing Condition of the White Crystal Sugar Agroindustry Supply Chain System.

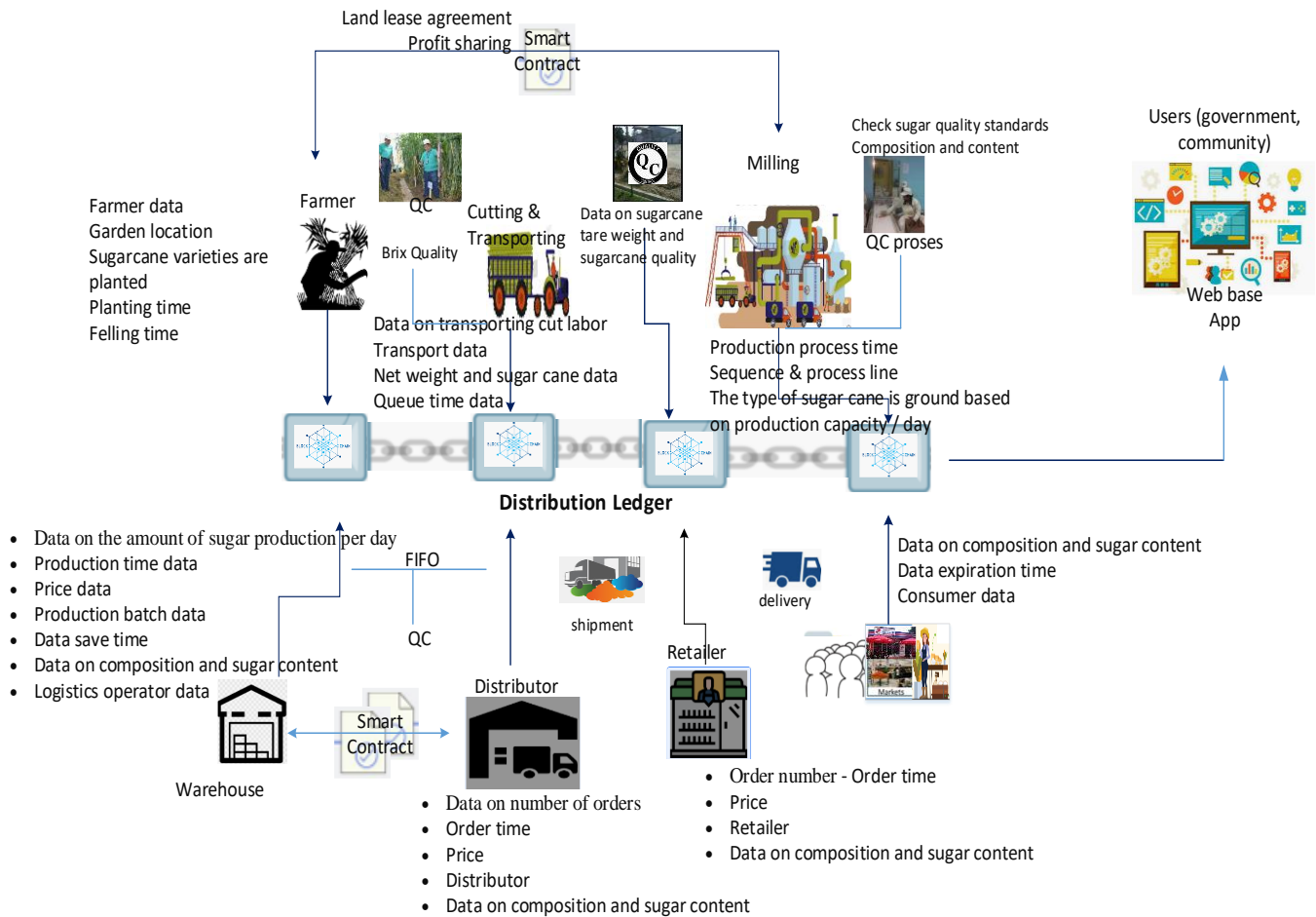


Fig. 4. Architectural Sugar Industrial Supply Chain based on Blockchain Technology.

How to solve traditional supply chain problems which are very sensitive and valuable for data trust and security, because in traditional and centralized systems we cannot prevent fraud

and errors in data information flowing from upstream to downstream or from the data of additional actors (government and financial institutions) which can make it difficult for

consumers to access and verify data sources[31]&[32]. Under the proposed title is the design of a blockchain-based white crystal sugar agro-industrial supply chain system, in which the blockchain can help enable unprecedented transparency, safely throughout the supply chain that will eliminate fraud and errors, improve logistics performance and supply chain management, minimize costs, reduce waste and process delays. According to the estimates from IBM [33], blockchain can increase global GDP by nearly 5% and total trading volume by 15%.

Blockchain is an unchanging and reliable shared ledger in storing transaction information of various data and actors in the supply chain. With a decentralized network and smart contracts, product information can be trusted, recorded securely, and does not change from upstream to downstream of the supply chain. The following is a proposal for designing a blockchain-based white crystal sugar agro-industry supply chain system.

Agroindustry supply chain which is globally inefficient, it is difficult to track problems that occur in the supply chain intentionally or unintentionally caused by corruption of data, product, and financial information that is not detected. The challenge of the supply chain now is to improve the data management functions that are still manual, centralized, data corruption, and data that are not real-time. Various efforts to improve supply chain performance become more efficient by increasing supply chain operations, to reduce the risks that occur.

Therefore, proposing the design of blockchain-based supply chain systems through the process of traceability and smart contracts is expected to improve the performance and competitiveness of the white crystal sugar agroindustry. So that it can benefit all actors in the supply chain and transparency and trust between actors can be increased as well.

Blockchain technology is expected to optimize supply chain performance, financing processes and accelerates process operations in the processing industry by providing safe, concise solutions, without using paper records and centralization. Can ease the manual process, ensure the accessibility, transparency, and integrity of the white crystal sugar agroindustry supply chain.

The blockchain network can also be applied in each of the supply chain actors so that it will form a reliable integrated information technology infrastructure and simplify workflow. Blockchain also facilitates accurate tracking of the quantity and quality of products along the chain to improve software performance, the performance of all actors in the supply chain network, which in turn can increase trust and transparency from the initial source until the product is consumed by the end consumers.

If these things are applied, the efficiency of supply chain performance will increase, intentional or unintentional mistakes will be reduced to zero defects and the benefits of actors along the chain can be maximized [34].

V. CONCLUSION

This research is expected to produce a design of a white crystal sugar agroindustry supply chain system based on blockchain technology. The previous traditional method which was still based on data centralization was changed to decentralized information, providing transparency and trust along with the supply chain actors with immutability based on cryptographic hash functions to secure data, tracking the truth of data quality, the quantity of inventory, and price, to provide product composition information. So that the blockchain-based traceability architecture design can be developed and implemented in white crystal sugar agroindustry by looking at the readiness of human, technological and financial resources. So that prototypes can be prepared based on the available blockchain consortium platforms such as Ethereum, by verifying and validating the supply chain system.

The suggestion raised is that this design can be made with a local platform that is starting to develop, such as the Vexanium Indonesia public blockchain, which has speeds above 2000 transactions per second and according to programmers is more suitable for retail use and users dealing directly with consumers such as social media activists.

ACKNOWLEDGMENT

This research received financial support from BUDI DN LPDP. I would like to thank the three supervisory commissions, IPB University who has provided knowledge, input, and suggestions for initial research on the sugar agro-industry, blockchain technology. I also want to thank my family (husband, children, and parents) who have supported my research.

REFERENCES

- [1] K. Pertanian, "Regulation of the Minister of Agriculture of the Republic of Indonesia No.68 / Permentan / OT.140 / 6/2013," 2013.
- [2] M. Asrol, M. Marimin, and M. Machfud, "Supply Chain Performance Measurement and Improvement for Sugarcane Agro-industry," *Int. J. Supply Chain Manag.*, vol. 6 No.3, no. September, pp. 8–21, 2017.
- [3] H. Suliantoro and D. Nugrahani, "Measurement and Evaluation of Supply Chain Performance Using the Balanced Scorecard-Analytical Network Process (BSC-ANP) Approach at PT.Madubaru Yogyakarta," in *SNST Proceeding*, 2015, pp. 17–23.
- [4] A. F. Fadhilah et al., "Performance Efficiency of Ant Sugar Supply Chain CV. Incised Politan in Kulon Progo Regency," *JoFSA*, vol. 1, no. 2, pp. 60–70, 2017.
- [5] A. M. Khushk, S. Sciences, A. Memon, and I. Saeed, "Analysis of sugar industry competitiveness in Pakistan," *J. Agric. Res.*, vol. 49(1), no. November, pp. 137–155, 2015.
- [6] D. J. Garcia and F. You, "Supply chain design and optimization: Challenges and opportunities," *Comput. Chem. Eng.*, vol. 81, pp. 153–170, 2015.
- [7] P. Stutterheim, "An Integrated Sugarcane Supply Chain Model: Development and Demonstration," University of KwaZulu-Natal, 2006.
- [8] J. Van Der Vorst, Performance measurement in agri-food supply-chain networks - An overview, no. January 2006. 2014.
- [9] M. Aravendan and R. Panneerselvam, "Literature review on network design problems in closed-loop and reverse supply chains," *Intell. Inf. Manag.*, no. May, pp. 104–117, 2014.
- [10] J. Jonkman, A. Kanellopoulos, and J. M. Bloemhof, "Designing an eco-efficient biomass-based supply chain using a multi-actor optimization model," *J. Clean. Prod.*, vol. 210, pp. 1065–1075, 2019.
- [11] A. Cheraghali-pour, M. M. Paydar, and M. Hajiaghaei-Keshteli, "Designing and solving a bi-level model for rice supply chain using the

- evolutionary algorithms,” *Comput. Electron. Agric.*, vol. 162, no. March, pp. 651–668, 2019.
- [12] N. Chiadamrong and R. Kawtummachai, “A methodology to support decision-making on sugar distribution for export channel: A case study of the Thai sugar industry,” *Comput. Electron. Agric.*, vol. 64, no. 2, pp. 248–261, 2008.
- [13] L. U. Opara, “Traceability in agriculture and food supply chain: A review of basic concepts, technological implications, and prospects,” *J. Food Agric. Environ.*, vol. 1, no. (1), pp. 101–106, 2003.
- [14] M. M. Aung and Y. S. Chang, “Traceability in a food supply chain: Safety and quality perspectives,” *Food Control*, vol. 39, pp. 172–184, 2014.
- [15] M. Fritz and G. Schiefer, “Tracking, tracing, and business process interests in food commodities: A multi-level decision complexity,” *Int. J. Prod. Econ.*, vol. 117, no. 2, pp. 317–329, 2009.
- [16] G. B. Zhang, Y. Ran, and X. L. Ren, “Study on product quality tracing technology in the supply chain,” *Comput. Ind. Eng.*, vol. 60, no. 4, pp. 863–871, 2011.
- [17] G. Barilla, A. Pinna, and G. Corrias, “Ensure Traceability in European Food Supply Chain by using a blockchain System,” in *WETSEB2019*, 2019, no. March, pp. 1–8.
- [18] S. Charlebois, M. Juhasz, L. Foti, and S. Chamberlain, “Food Fraud and Risk Perception: Awareness in Canada and Projected Trust on risk-mitigating Agents,” *J. Int. Food Agribus. Mark.*, vol. 29, no. 3, pp. 260–277, 2017.
- [19] M. V. Kumar and N. C. S. N. Iyengar, “A Framework for Blockchain Technology in Rice Supply Chain Management Plantation,” in *Advanced Science and Technology Letters*, 2018, no. vol 146, pp. 125–130.
- [20] P. Helo and A. H. M. Shamsuzzoha, “Real-time supply chain—A blockchain architecture for project deliveries,” *Robot. Comput. Integer. Manuf.*, vol. 63, pp. 1–14, 2020.
- [21] R. Casado-vara et al., “How blockchain improves supply the supply chain : case study alimentary chain supply chain supply chain,” *Procedia Comput. Sci.*, vol. 134, pp. 393–398, 2018.
- [22] G. Zhao et al., “Blockchain technology in agri-food value chain management: A synthesis of applications, challenges, and future research directions,” *Comput. Ind.*, vol. 109, pp. 83–99, 2019.
- [23] V. Thiruchelvam, A. S. Mughisha, M. Shahpasand, and M. Bamiah, “Blockchain-based Technology in the Coffee Supply Chain Trade : Case of Burundi Coffee,” *J. Telecommun. Electron. Comput. Eng.*, vol. 10, no. 3–2, pp. 121–125, 2018.
- [24] K. Salah, N. Nizamuddin, R. Jayaraman, and M. Omar, “Blockchain-Based Soybean Traceability in Agricultural Supply Chain,” *IEEE Access*, vol. 7, pp. 73295–73305, 2019.
- [25] G. Perboli, S. Musso, and M. Rosano, “Blockchain in Logistics and Supply Chain: A Lean Approach for Designing Real-World Use Cases,” *IEEE Access*, vol. XX, pp. 1–12, 2018.
- [26] A. A. Arsyad, S. Dhadkah, and M. Koppen, “Two-Factor Blockchain for Traceability Cacao Supply Chain,” *springer Nat. Switz. Ag*, vol. 23, pp. 332–339, 2019.
- [27] F. Tian, “An agri-food supply chain traceability system for China based on RFID & blockchain technology,” in *13th International Conference on Service Systems and Service Management*, 2016, pp. 1–6.
- [28] J. F. Galvez, J. C. Mejuto, and J. Simal-Gandara, “Future challenges on the use of blockchain for food traceability analysis,” *Trends Anal. Chem.*, pp. 1–43, 2018.
- [29] H. Feng, X. Wang, Y. Duan, J. Zhang, and X. Zhang, “Applying blockchain technology to improve agri-food traceability: A review of development methods, benefits, and challenges,” *J. Clean. Prod.*, vol. 260, 2020.
- [30] W. A. Kusumo, Y. Setiowati, and K. Fathoni, “The design of information system for sugar cane milling in a sugar company as a case study of the new pesantren sugar factory-Kediri,” 2013.
- [31] S. Abeyratne and R. . Monfared, “Blockchain ready manufacturing supply chain using a distributed ledger,” *Int. J. Res. Eng. Technol.*, vol. 05, no. 09, pp. 1–10, 2016.
- [32] P. Helo and Y. Hao, “Blockchains in operations and supply chains: A model and reference implementation,” *Comput. Ind. Eng.*, vol. 136, no. July, pp. 242–251, 2019.
- [33] M. H. Ronaghi, “A blockchain maturity model in the agricultural supply chain,” *Inf. Process. Agric.*, pp. 1–10, 2020.
- [34] E. Chen, “An approach for Improving Transparency and Traceability of Industrial Supply Chain with Blockchain Technology,” *Tampere University of Technology*, 2016.

Fraud Detection in Shipping Industry using K-NN Algorithm

Ganesan Subramaniam¹

College of Computing and Informatics
Universiti Tenaga Nasional
Malaysia

Moamin A. Mahmoud²

Institute of Informatics and Computing in Energy (IICE)
Universiti Tenaga Nasional
Malaysia

Abstract—The shipment industry is going through tremendous growth in volume thanks to technological innovation in e-commerce and global trade liberalization. Volume growth also means a rise in fraud cases involving smuggling and false declaration of shipments. Shipping companies and customs are mostly relying on routine random inspection thus finding fraud is often by chance. As the volume increases dramatically it would no longer be sustainable and effective for both shipment companies and customs to pursue traditional fraud detection strategies. Other related papers on this area have proven that intelligent data-driven fraud detection is proven to be far more effective than routine inspections. However, the challenge in data-driven detection is its effectiveness are often reliant on the availability of data and the various fraud mechanism used by fraudsters to commit shipment related fraud. As such in this paper, we review and subsequently identify the most optimized approaches and algorithms to detect fraud effectively within the shipping industry. We also identify factors that influence fraud activity, review existing fraud detection models, develop the detection framework and implement the framework using the Rapidminer tool.

Keywords—*Fraud detection; shipping industry; K-NN algorithm*

I. INTRODUCTION

World Customs Organization's (WCO) Illicit Trade Report 2016 states that the year 2016 was marked as the year of Digital Customs where the administrations were encouraged to actively showcase and promote the use of Information and Communication Technologies (ICT) by using a data-driven approach to collect and safeguard duties, control the flow of goods, people finally to secure cross-border trade [1]. CIO Magazine from IDG identified fraud detection as one of the IT projects primed for machine learning [2]. As such, there is a pressing need within the industry for a more intelligent fraud detection system that can considerably improve the detection of wrong declarations and smuggling compared to random checks.

Liberalization in trade and technological innovation such as advancement in e-commerce has accelerated international shipping volumes significantly over the last few years. The latest statistics from The International Air Transport Association (IATA) released full-year 2017 data for global air freight markets showing that demand, measured in freight ton-kilometers (FTKs) grew by 9.0%. This was more than double the 3.6% annual growth recorded in 2016 [3]. Such a rapid rise

in volumes will lead to an increase in safety and compliance issues as tremendous volume creates a strain for both the shipping companies and customs authorities to perform safety and compliance audits on most shipments. Meanwhile, on the end of the spectrum, customers are demanding e-commerce providers and shipping companies to provide faster deliveries. By fulfilling urgent deliveries shipping companies are also able to expand their business to a new market segment such as urgent medical deliveries which provides a higher margin of profit to shipping companies. Thus, placing more manual checks on shipments by both shipping companies and customs will only cause more delays on shipments which directly impact the profitability of shipping companies and their customer base. As such, there is an urgent need to automate and increase the effectiveness of the current random checking done by both the shipping companies and the customs. There are various violations committed by shippers such as illicit trade, smuggling that violates shipping restrictions between countries or miscoding of the shipment items shipped that saves custom duties payments. There could be also security and safety issues if shipments are not audited thoroughly. Imagine the impact of dangerous or flammable goods such as mobile batteries being declared as safe goods in an air freight shipment that can cause devastating impacts such as plane crashes. According to the UK P&I Club, 27% of incidents on cargo ships in 2013 and 2014 were attributable to mis-declared hazardous cargo, second only to poor packaging [3].

Issues and challenges highlighted in [4] [5] paper are generally concept drift or dynamic fraud patterns, overlapping data, capability to support real-time detection requirements, skewed distribution, integrating a vast amount of data, and data quality-related issues. Concept drift is the challenge to deal with sudden customer behavioral changes which could turn out to be a false positive outcome. The solution to overcome concept drift is to use an adaptive FDS algorithm that learns and improves over time by factoring in all the possible input variables that may influence the change in expected behaviors. Overlapping data is the issue where fraudulent transactions are made to look like genuine data which becomes true negative cases. The skewed distribution is the issue of having a very low ratio of fraudulent cases which may not be sufficient to train supervised classification-based FDS algorithms. Data quality issues also need to be reviewed as this factor directly impacts the efficiency of fraud detection.

Each issue and challenge impacts the respective fraud domain area differently. There are mainly five business domain

areas where FDS has applied namely banking, telecommunication, insurance, online business, and shipping [6]. The specific area involved in banking is credit card-related fraud. Meanwhile, in insurance, the specific area will be medical insurance claims and vehicle insurance claim-related fraud [7]. In an online business, the typical fraud area is online auction-related fraud. Shipping-related fraud is usually related to smuggling and miscoding which is a false declaration of the goods being shipped. The critical issue for the shipment domain is getting efficient real-time results within a huge data set. The bigger the data set the better the efficiency thus we need to have a solution architecture that can process an optimal volume of the desired dataset that is efficient enough to be executed in a real-time mode. In the express shipping domain, accuracy and real-time performance is very critical as the life cycle of a shipment only varies between 3-5 days depending on the weight and location. Thus, identifying the fraud before the shipment gets delivered is very critical. The earlier the shipments are intercepted the bigger the cost benefits for both organizations and customers. Immediate detection avoids revenue leakage and improves customer's trust and confidence towards the organization's brand. It also ensures fraud culprits are identified effectively and handed over much earlier to authorities that may help to reduce future fraud cases.

To build a solution that detects fraud effectively we need to identify the parameters or the data elements that influence the most in actual fraud cases. In a study done in a leading global logistics company, it was identified that location is one of the key parameters that influence fraud cases. The location for shipment can be either the origin or the destination of the shipment. According to a report published by World Customs Organization (WCO), shipment origin and destination location can be a major factor as most frauds tend to originate from or being sent to a specific location. Based on data provided by a major global logistics organization the data that will be extracted for our simulation will be the origin and destination respective latitude and longitude values. Since these values are numerical it can precisely identify a location. With these precise numerical-based location values, a specific fraudulent shipment origin or destination and its surrounding area within a defined radius will be tagged as fraudulent by the algorithm. By having numerical data, the processing speed of the algorithm will also be much faster as opposed to using text or image data.

II. LITERATURE REVIEW

In subsequent sections, we will be reviewing various papers that are related to fraud detection.

A. Methods

The search strategy is the definition and selection process to find the most relevant papers are described in the following. The digital databases searched in this review include IEEE Xplore, Springer, and Science Direct. The reason for the selection of these four databases is due to the availability of highly cited and reliable papers in the fields of computer science and its related applications. The review objective is to find all primary research work associated with fraud detection systems within the shipping or logistics domain. The earlier phrase that was searched is "fraud detection system and

shipping or cargo or freight or logistics" but since there were not many fraud detection system papers in the shipping domain thus most of the returns were only relevant to fraud detection. There were only three papers related to the shipping domain [8]. Finally, only the term "fraud detection system" was used. The initial query resulted in a total of 5866 papers: 598 from IEEE Explore, 964 from Science Direct, and 4304 from Springer. The filtered articles were published between 2000 and 2018. For Science Direct and Springer besides the year filter based on the topic is also applied to ensure non-computer science-related papers are excluded. The reason the year was narrowed down between 2000 and 2018 was due to the no of results which came up to thousands. After sifting through some of the papers we have divided into survey papers and specialized papers which specifically delve into specific techniques or business domain area such as financial which is a credit card or insurance, healthcare, telecommunication, and internet-related marketing fraud.

B. Review

The earliest survey paper since the year 2000 is the paper from [9]. This paper reviews fraud detection from a statistical perspective. Just like most fraud detection-related papers this paper also categorizes basic statistics models for fraud detection methods into supervised and unsupervised. Besides categorization by models, it also surveys papers based on application area or domain. Among the application covered are in the area of credit card fraud, money laundering, telecommunications fraud, computer intrusion which is also known as hacking these days, medical and scientific fraud which also includes plagiarism in the education sector. This paper concluded that the key issue in fraud detection is the effectiveness of fraud detection. Factors such as the speed of detection are directly related to its effectiveness. As such, a strategy to use a graded system of investigation is suggested where areas with very high suspicion and high fraud value merit immediate and intensive investigation. This paper also concluded that fraud detection can be achieved even in difficult circumstances but there are also many challenges and opportunities waiting to be tapped in the future. In 2004 another fraud detection survey paper by [10] was published. This paper focuses more on fraud detection techniques. Domain areas covered are credit card fraud detection, telecommunication fraud detection, and computer intrusion detection. Common techniques applied in credit cards are outlier detection which is an unsupervised method that does not rely on historical data. Outliers are based on observation of deviation against the normal or average pattern. It's suitable to detect fraud that has not previously occurred. To detect fraud pattern which previously occurred then supervised method using historical labeled data are used. Neural network-related techniques which are a set of interconnected weighted nodes designed to function like a human brain are also applied widely for credit card fraud detection, but this technique requires an actual data set that is rarely made available to the public. For computer intrusion detection several techniques such as expert system, neural networks, model-based reasoning, data mining, and state transition analysis are applied. The challenge in the computer intrusion domain is to deal with heaps of the audit trail data, dealing with false alarms rate, difficulty in testing, simulating potential scenarios, and poor portability as the

ruleset is very specific to a particular environment. Lastly in telecommunication fraud detection among the techniques used are rule-based, a neural network that includes Bayesian network and also visualization methods. The challenge of managing the data load in supervised learning for the rule-based and neural network can be mitigated by using unsupervised learning to filter out normal behavior data. To create a more robust selection process for rule base technique a non-greedy rule-selection approach can be explored further. The telecommunication environment is very dynamic and always evolving thus it requires accurate definitions of thresholds and parameters that in tune with the changing landscape of this domain.

In 2010 data mining-based fraud detection research surveyed papers from the year 1998 till 2010 [11]. This paper also highlights types of fraudsters and affected industries. The type of fraudsters is divided broadly into managers, employees, or external parties. The most challenging fraudsters are the external parties as they are many of them and they can make use of various complex and new fraud mechanisms. This is the area where we need to apply data analytics or data mining techniques as it will be cost-effective compared to conventional manual methods to find the riskiest parties by using suspicion scores, rules, and visual anomalies that can be investigated and refined. This paper also identifies the fraud domain area as internal which is fraud committed by management and staff within the organization, insurance, credit card, and telecommunications. Credit transactional fraud detection has received the most attention from researchers. There are also other emerging fraud areas such as e-business and e-commerce related fraud in the online world. Two main challenges of data mining-based fraud detection research are the lack of publicly available real data to perform research on and also the lack of well-researched methods and techniques. To overcome the challenge of data availability a solution to use simulated data that closely matches the actual data which are often very sensitive to be shared in public domains. These were proposed in some papers such as [12] [13] [14] [15]. To overcome the issue well-researched methods and techniques some performance matrices and measures are critical to ensure fraud detection gets well-deserved attention from business stakeholders to invest and provide funding that flourishes research and development in these initiatives. Among the measures taken are such as placing a monetary value on predictions that can maximize cost savings/profits by having their own cost and benefit model customized according to their respective business needs. Other considerations to determine the methods of fraud detection are speed of fraud detection and

also the styles/types of detection such as online/real-time or batch mode.

In early 2016, Abdallah et al. [4] released a survey paper that covers papers from 1997 to 2014. This paper provides a good summary of the matters surrounding the fraud detection system. Fraud is defined by the Association of Fraud Examiners (ACFE) as the use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resource or assets. There are 2 main types of fraud systems namely fraud prevention systems and fraud detection systems. Fraud prevention is the first line of defense against fraud which blocks the entry of any fraudsters into the system. Meanwhile, fraud detection is the next layer of defense that detects fraudsters who have already committed the fraudulence act. Over the years many fraud detection approaches and techniques have been applied.

In [16], the paper studied the highest level grouping been categorized by area of study. The 2 main groups are statistical modeling and machine learning. Statistical modeling is an area of mathematics that deals with collecting and analyzing data with some assumptions. Machine learning is a technique using programming algorithm models that learn from data and solve complex problems. There are 2 main methods of machine learning namely supervised and unsupervised types. Some approaches combine these two techniques which are known as semi-supervised. The difference between supervised and unsupervised is in the use of labeled data in supervised as oppose to unsupervised which does not use any labeled data. Labeled data is the identification of fraud data in the data set that are used to train the algorithm or model. Unsupervised techniques rely on a grouping of similar attributes or finding outliers that can identify unusual behavior or patterns that can be further investigated. An overview of the various methods and techniques is illustrated in Fig. 1.

Table I provides a summary of fraud detection data mining tasks with commonly used algorithmic techniques and example use cases [17]. Table II provides a comparison summary of fraud detection data mining algorithms that would help to identify the suitable algorithm that can be applied in this paper's use case [17].

Another potential mechanism that could be used in Fraud detection is using multi-agent systems [19] [26] [27]. MAS could be integrated with social norms [18] [20] [21] [22], to identify and learn different customs norms and accordingly predict anomaly behaviour [23] [24] [25] [28].












Fig. 1. Fraud Detection Algorithms.

TABLE I. FRAUD TASKS SUMMARY

Tasks	Description	Algorithms	Approach	General Usage
Classification	Datapoint prediction within predefined groups. Learning-based prediction from known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors	Supervised	Classifying customers into known groups. e.g., Profitable customers/fraudsters
Regression	Numeric target label prediction of a data point. Learning-based prediction from known data set.	Logistic regression	Supervised	Subsequent period fraud predicting fraud and estimating losses
Anomaly detection	Outlier prediction of datapoint against other data within the data set.	Distance-based, density based, local outlier factor (LOF)	Unsupervised	Detection of Credit Card Fraud and network intrusion
Clustering	Natural clusters identification based on inherent properties within data sets.	k-means, density-based clustering	Unsupervised	Fraud segments identification within organization using transaction, web, and customer call data

TABLE II. ALGORITHMS USED FOR FRAUD DETECTION COMPARISON SUMMARY TABLE

Classification						
Algorithm	Model	Data	Outcome	Strength	Weakness	Operational Usage
Decision Trees  Branching out data into subsets where each contains responses of one class	Data set partitioning based on different predictors values	Unrestricted variable type for predictors	The label must be categorical.	Simpler to present to business users. Predictors normalization is not required	Data overfit. Input data changes can cause significantly different trees. Challenging to choose the parameter	Marketing segmentation, fraud detection
k-Nearest Neighbors  Lazy learner where no model is defined. New unknown data point is compared with similar known training set data point	Model is the entire training data set.	Unrestricted but distance calculations work better with numeric data. Normalized data required.	Target variable prediction, which is categorical.	Faster to build model. Missing attributes handled well Works with nonlinear relationships.	Operational runtime and storage requirement will be high. Value of k randomly selected. No model description	Image processing, application, fraud detections
Naïve Bayesian  Bayes theorem output class prediction by calculating class conditional probability and prior probability.	For each attribute with an output class need a lookup table of probabilities and conditional probabilities	Unrestricted but calculation of probability works better with categorical attributes	Probability prediction for all class values, together with the winning class.	Faster modeling and deployment. Suitable algorithm for benchmarking. Strong statistical foundation	Training data set needs represent population sample well and needs to have complete input/output combination. Independent attributes required	Detections of Spam and mining of text.
Artificial Neural Networks  Biological nervous a system inspired mathematical model. Actual /prediction tuning using network weights	Processing of data based on layers of network topology and weights.	All attributes should be numeric.	Prediction of target (label) variable, which is categorical	Suitable for modeling nonlinear relationships. Fast response time during runtime.	Complex inner working of the model. Requires preprocessing of data. Missing attributes cannot be handled.	Image recognition, fraud detection, quick response time applications.
Regression						
Algorithm	Model	Data	Outcome	Strength	Weakness	Operational Usage
Support Vector Machines  Boundary detection algorithm that illustrates multidimensional boundaries separating data points that belong to different classes.	Vector equation model that enables classification of new data points into different regions or groups.	All attributes should be numeric.	Prediction of target (label) variable, which can be categorical or numeric.	Underfit data and tolerates high variance. Small changes to input data does not influence boundary that results from inconsistency. Suitable for nonlinear relationships.	Training phase computational performance is slower. Additional effort is also needed to optimize parameter combinations.	Optical character recognition, fraud detection, modeling unpredictable events.

Clustering						
Algorithm	Model	Data	Outcome	Strength	Weakness	Operational Usage
k-means  Finding k centroids once data set is divided into k clusters or groups	Find k centroids using algorithm and data points are associated to the nearest centroids, that forms a cluster or group.	Data should be normalized. Works with all types of data but for distance calculations works better with numeric data.	Data set is appended by one of k cluster labels.	Simple implementation. Can be used to reduce number of dimension.	K specification may not be precise and Cluster may not be natural clusters. Sensitive to outliers.	Customer segmentation, anomaly detection, Applicable for natural globular clustering.
Anomaly Detection						
Algorithm	Model	Data	Outcome	Strength	Weakness	Operational Usage
Distance Based  Outlier Identification based from kth nearest neighbor	Distance score assigned for all data point based on nearest neighbor.	Data must be normalized due to distance calculation Numeric and categorical attributes accepted.	Distance score assigned to every data point. The further the distance, higher the probability of an outlier.	Easier implementation. Suitable for numeric attributes	Specification of k is arbitrary.	Fraud detection, pre-processing technique.
Density Based  Identification of outlier based on data points in low-density regions.	Neighborhood based density score for all data points	Data must be normalized due to density calculation. Both numeric and categorical attributes accepted.	Density score assigned to every data points. The lower the density, the probability of an outlier is higher.	Easier Implementation. Higher precision with numeric attributes.	Distance parameter specification by the end user. Challenge in identifying varying density regions.	Fraud detection, preprocessing technique.
Local outlier factor  Outlier identification based on relative density calculation within neighborhood	Neighborhood based relative density score for all data points.	Data must be normalized due to density calculation. Both numeric and categorical attributes accepted.	Density the score assigned to every datapoints. The lower the relative density, the probability of an outlier is higher.	Handles varied density scenario.	Distance parameter specification by the end user.	Fraud detection, preprocessing technique.

III. SOLUTION DESIGN

The processes defined for the proposed model as shown in Table III. First, the fraud data will be prepared, normalized, and cleaned up by replacing missing values and removing duplicates. The data is simulated based on a study done in a global logistics organization based on their historical shipment origin and destination data for five years from 2012 till 2017. There are 5 columns namely Fraud label, Origin City Latitude, Origin City Longitude, Destination City Latitude, and Destination City Longitude. Origin or destination with high cases of fraud is tagged with fraud field as “Y”. The number of rows simulated is 1500 records. A snapshot of the data is shown in Table III.

Once the data preparation is completed data will be fed into the process model. Fraud attribute shall be labeled as target role. Once the label has been set up the validation process can be configured. Split type can be set as relative, and the ratio can be dynamically changed from a range of 0.6 to as 0.8 to increase the accuracy. A split ratio of 0.75 means 75% of the data will be used as training the algorithm and the remaining 25% data will be used to test the trained algorithm. Model design is illustrated in Fig. 2. The flow in blue is using the labeled data are iterated with various combinations of the split ratio, various algorithms, and parameters related to the specific algorithm. Tuning of these variables will produce results that can be measured in terms of accuracy within an acceptable execution time.

Once an acceptable performance is achieved the algorithm chosen together with its known parameters can be applied to

every new incoming shipment data. In this way, fraudulent shipments can be detected at the time the shipment is still in progress within the network before it gets delivered. The algorithm will be updating the prediction column to flag fraudulent shipments accordingly. Shipments flagged as fraud can be further investigated to check if it is genuine. If it's wrongly flagged further analysis needs to be done to improve the algorithm in the future. The analysis also needs to be done on shipments that were not flagged by the algorithm, but it was later found to be fraudulent.

TABLE III. DATA SET

Fraud	Origin City Lat	Origin City Lng	Dest City Lat	Dest City Lng
Y	-28.5495	29.78	7.8704	9.78
N	38.0517	58.21	40.306	36.563
N	13.979	45.574	30.5333	105.5333
N	43.8436	-88.8386	-25.5096	-57.36
N	48.5095	-122.2344	35.2495	-81.1856
N	42.7666	-78.6172	38.758	-89.9839
N	10.9587	123.3086	23.1904	75.79
Y	-13.6396	-72.89	44.5372	135.5172
N	-33.5995	150.74	72.685	-78.0001
Y	40.8128	44.4883	-0.215	-78.5001

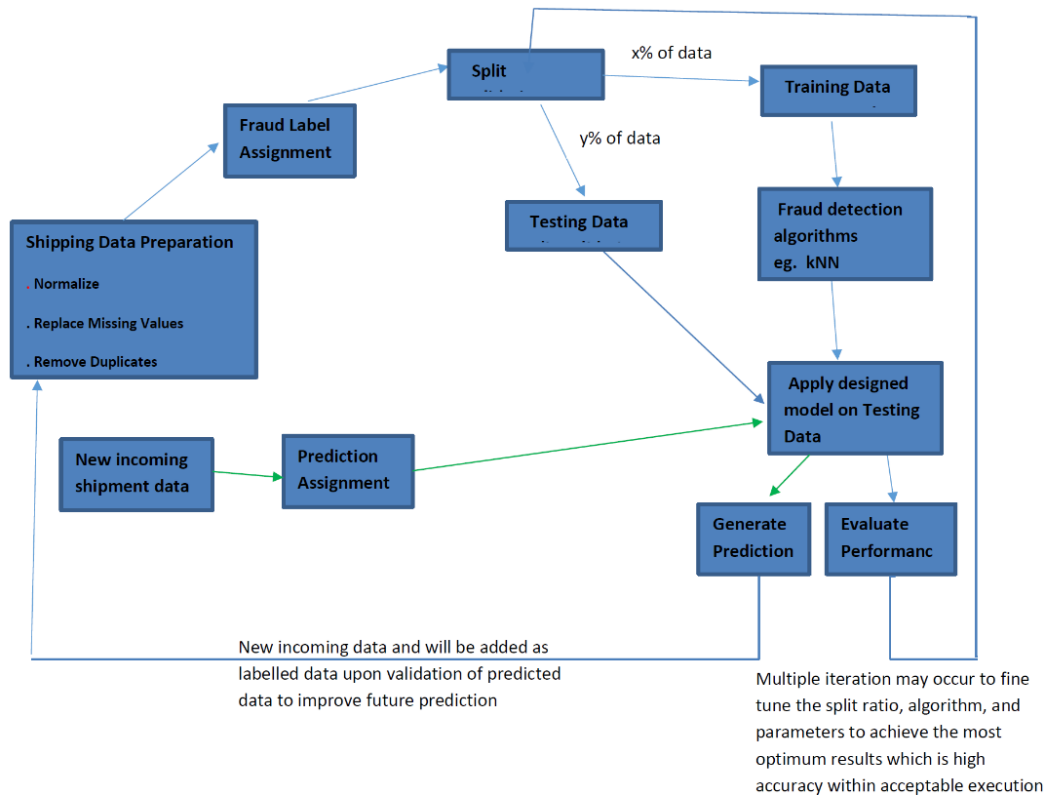


Fig. 2. Solution Design.

IV. SOLUTION MODELING

Various tools can be used to perform fraud detections in the market like R, Rapid-Miner, SAS Enterprise Miner, IBM SPSS, etc. In this paper, we have selected Rapidminer as it is simple to use and it also has many built-in ready-to-use algorithms. This allows various techniques can be tested against the available data. Furthermore, it's also provided as a freeware version for students. Once the data is ready then the Rapidminer tool can be used to set up the model. Data used for this modeling will be as per below.

Among the algorithm tested are Naive Bayes, Neural Net, Deep Learning, Decision Tree, Logistic Regression, SVM and finally k-Nearest Neighbors or k-NN as shown in Fig. 3.

After several executions with various algorithms and split ratio combination, it was found that the optimal best result with the highest accuracy of about 98.4% was achieved using the k-NN algorithm using default parameters as shown in Fig. 4.

As shown in Fig. 5, in terms of execution speed it's found that most of the algorithm immediately returned the result except for Neural Net that took almost 2 seconds, and Deep Learning that took 6 seconds. As such these two algorithms are not suitable for fraud detection within the shipping domain as speed is one of the key criteria.

The above figure represents the relationship between the k-NN key parameter which is the k nearest neighbor number of classes against the accuracy of the prediction. It's was found that the highest accuracy was recorded when k is either 1 or 2. Accuracy starts to drop once k is increased beyond 2.

Fig. 6 represents the relationship between the k-NN key parameter which is the k nearest neighbor number of classes against the accuracy of the prediction. It's was found that the highest accuracy was recorded when k is either 1 or 2. Accuracy starts to drop once k is increased beyond 2. In this study, the k-NN algorithm has been identified as the best optimum results in terms of accuracy and speed criteria that were required in fraud detection within the shipping domain. Results in Fig. 6 illustrates that genuine fraud was detected correctly for 88% of the total cases. Non-fraud cases were predicted correctly at 99.14% of the total cases.

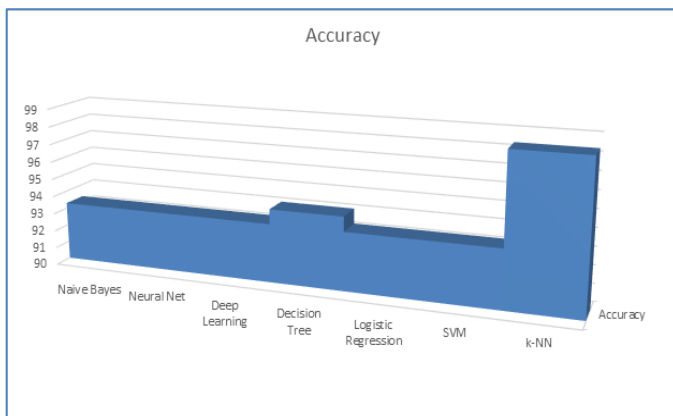


Fig. 3. Accuracy Results According to Algorithm Type.

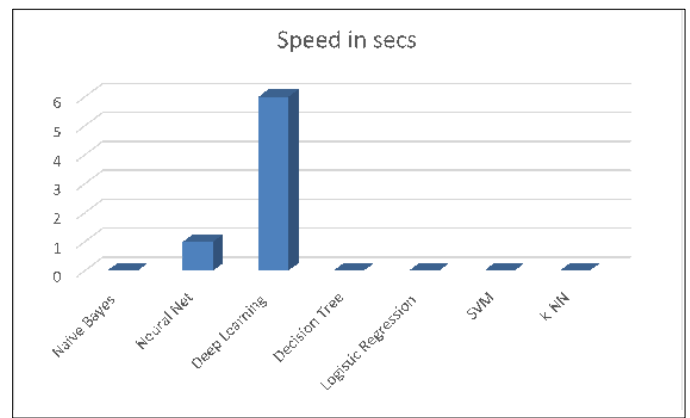


Fig. 4. Execution Speed According to Algorithm Type.

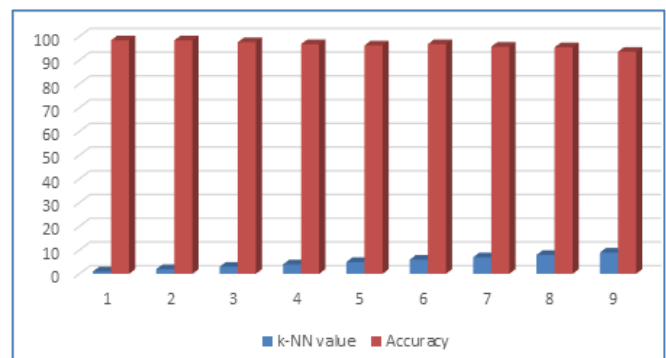


Fig. 5. Results According to different Values of k within the k-NN Algorithm.

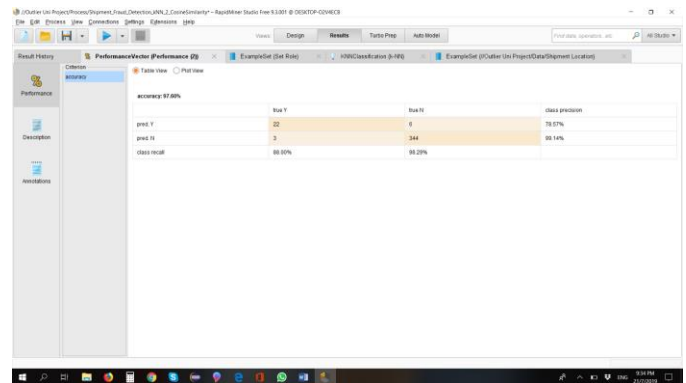


Fig. 6. Results.

As such to achieve high accuracy with optimum performance k-Nearest Neighbour algorithm technique is proposed to detect fraud in the shortest possible time within the shipping domain. This technique is proposed based on the modeling simulation done in Rapidminer. This algorithm indicates that this technique usually provides an acceptable response time during execution which is within a second. Thus, the model for this solution will be as shown in Fig. 7 below where the shipping data will be first pre-processed to ensure there are no missing values. The pre-processed data will be then split between training and test data. The data set available can be split between training and test data with 75% for training the algorithm and 25% for testing the algorithm which is also close to the split recommended by [17].

Performance evaluation with new incoming data and subsequent execution is with more historical data. New unlabeled data can be routed to the model to predict if it could be fraudulent as shown below in 7e 7. The set of new data will be analyzed by the trained algorithm and will predict each row of data with origin longitude latitude and destination longitude-latitude with a prediction flag to each row. Suspected fraudulent locations will be tagged as Y and non-fraudulent will be identified as "N". The identification is primarily based on how close the locations to the labeled fraudulent cases are provided in the learning stage. Thus, if new locations are present in the data these data will not be identified as a fraud as there is no historical data linked to it. To resolve this challenge new location data can be identified in outlier techniques and analyzed distinctly before it's can be used as part of the main dataset.

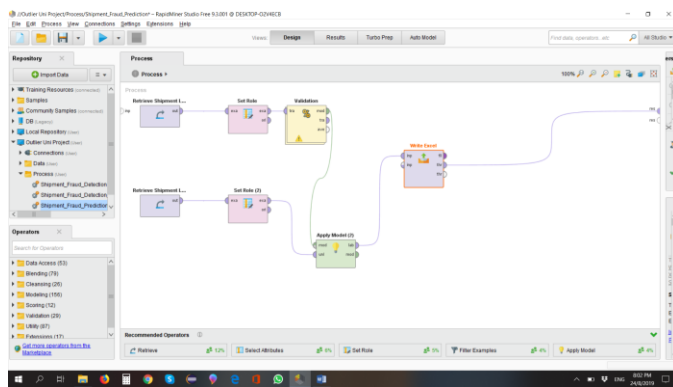


Fig. 7. New Data Test Model.

Alternatively, if the customer data profile is available K-NN distance-based outlier techniques can be applied by looking for any outlier locations within a customer's shipment data set. If there are no new locations expected from a particular customer and if there are any new locations detected within the customer's data set then this data can be classified as potential fraud using outlier techniques. As such combining machine learning, K-NN and outlier techniques can be a complementary strategy to increase the effectiveness of fraud detection in shipping domains.

V. CONCLUSION

As a conclusion, we are recommending K-NN algorithm machine learning to address fraud detection within the shipping domain. It's proven that fraud detection using machine learning is much more efficient compared to manually identifying fraud while a shipment is in progress. Identifying fraud before the shipment arrives at the destination is very crucial to ensure that fraud items do not get delivered to the consignee. This is only possible by automating the fraud detection process as some international shipments can be delivered within the same day depending on the location. Using the K-NN algorithm ensures fraud can be detected within the duration of shipment which is an effective way to stop the current fraud and to reduce future fraud cases. To overcome some challenges in identifying new cases of fraud K-NN machine learning technique can be combined with distance-based outlier techniques on data set that are grouped by customer profiles. Getting hold of actual shipping data from logistic companies was quite challenging

due to the sensitive nature of shipping data. As such exploration of the various detection approaches, analyzing the strength and weakness of each before choosing the most optimum approach was done with simulated data which was based on parameters identified in a study done in a shipping company. Future papers may use these approaches and algorithms from this paper to simulate and perform further testing with actual data if they have access to it. Besides location parameters which were used in this paper, other parameters influence fraud in the shipping domain such as shipment weight, payment method, and the profile of the customer which was not evaluated in this paper due to lack of actual production data. As such in future papers these parameters can be considered to get results with higher accuracy.

ACKNOWLEDGMENT

This work is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No. J510050002.

REFERENCES

- [1] Shelley, L. I. (2018). *Dark commerce: How a new illicit economy is threatening our future*. Princeton University Press.
- [2] Horsey, L. L. (2017). *Data Analytics for Fraud Prevention and Detection in State Government* (Doctoral dissertation, Utica College).
- [3] International Air transport Association(IATA), 2018. <http://www.iata.org/pressroom/pr/Pages/2018-01-31-01.aspx>.
- [4] Abdallah A., Maarof M. A., Zainal A., Fraud detection system: A survey, *Journal of Network and Computer Applications*, Volume 68, 2016, Pages 90-113,.
- [5] S. Makki et al., "Fraud Analysis Approaches in the Age of Big Data - A Review of State of the Art," 2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Tucson, AZ, USA, 2017, pp. 243-250, doi: 10.1109/FAS-W.2017.154.
- [6] Panigrahi, S., Kundu, A., Sural, S., & Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning. *Information Fusion*, 10(4), 354-363.
- [7] Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105, 87-95.
- [8] Triepels, R., Daniels, H., & Feelders, A. (2018). Data-driven fraud detection in international shipping. *Expert Systems with Applications*, 99, 193-202.
- [9] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical science*, 17(3), 235-255.
- [10] Kou, Y., Lu, C. T., Sirwongwattana, S., & Huang, Y. P. (2004, March). Survey of fraud detection techniques. In *IEEE International Conference on Networking, Sensing and Control, 2004* (Vol. 2, pp. 749-754). IEEE.
- [11] Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
- [12] Barse, E., Kvarnstrom, H. & Jonsson, E., "Synthesizing Test Data for Fraud Detection Systems", *Proc. of the 19th Annual Computer Security Applications Conference*, 384-395, 2003.
- [13] Chen, R., Chiu, M., Huang, Y. & Chen, L. , "Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines", *Proc. of IDEAL2004*,800-806 , 2004.
- [14] Aleskerov, E., Freisleben, B. & Rao, B. , "CARDWATCH:A Neural Network-Based Database Mining System for Credit11 Card Fraud Detection" , *Proc. of the IEEE/IAFE on Computational Intelligence for Financial Engineering*, 220-226 , 1997.
- [15] Pathak, J., Vidyarthi, N. & Summers, S. , "A Fuzzy-based Algorithm for Auditors to Detect Element of Fraud in Settled Insurance Claims", *Odette School of Business Administration* , 2003.

- [16] Richard J. Bolton and David J. Hand , “Statistical Fraud Detection: A Review” , Imperial College, 2002.
- [17] Vijay K , Bala D , “Predictive Analytics and Data Mining“ , Elsevier Inc , 2015.
- [18] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., & Mustapha, A. (2011, July). Norms detection and assimilation in multi-agent systems: a conceptual approach. In Knowledge Technology Week (pp. 226-233). Springer, Berlin, Heidelberg.
- [19] Jassim, O. A., Mahmoud, M. A., & Ahmad, M. S. (2015). A multi-agent framework for research supervision management. In Distributed Computing and Artificial Intelligence, 12th International Conference (pp. 129-136). Springer, Cham.
- [20] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., Mustapha, A., & Hamid, N. H. A. (2013, May). Obligation and Prohibition Norms Mining Algorithm for Normative Multi-agent Systems. In KES-AMSTA (pp. 115-124).
- [21] Mahmoud, M. A., Ahmad, M. S., Ahmad, A., Mustapha, A., Yusoff, M. Z. M., & Hamid, N. H. A. (2013). Building norms-adaptable agents from potential norms detection technique (PNDT). International Journal of Intelligent Information Technologies (IJIT), 9(3), 38-60.
- [22] Mahmoud, M. A., Mustapha, A., Ahmad, M. S., Ahmad, A., Yusoff, M. Z. M., & Hamid, N. H. A. (2013). Potential norms detection in social agent societies. In Distributed Computing and Artificial Intelligence (pp. 419-428). Springer, Cham.
- [23] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Mostafa, S. A. (2018, February). A regulative norms mining algorithm for complex adaptive system. In International Conference on Soft Computing and Data Mining (pp. 213-224). Springer, Cham.
- [24] Mahmoud, M., Ahmad, M. S., Mostafa, S., & Subramainan, L. (2020). How Norm Assimilation Using Agent-Based Systems. Journal of Systems Science and Complexity, 33(4), 849-881.
- [25] Mahmoud, M. A., Ahmad, M. S., & Mostafa, S. A. (2019). Norm-based behavior regulating technique for multi-agent in complex adaptive systems. IEEE Access, 7, 126662-126678.
- [26] Mahmoud, M. A., & Ahmad, M. S. (2016, August). A prototype for context identification of scientific papers via agent-based text mining. In 2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR) (pp. 40-44). IEEE.
- [27] Mahmoud, M. A., & Ahmad, M. S. (2015, August). A self-adaptive customer-oriented framework for intelligent strategic marketing: A multi-agent system approach to website development for learning institutions. In 2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR) (pp. 1-5). IEEE.
- [28] Mahmoud, M. A., Ahmad, M. S., Yusoff, M. Z. M., & Mustapha, A. (2014, December). Norms assimilation in heterogeneous agent community. In International Conference on Principles and Practice of Multi-Agent Systems (pp. 311-318). Springer, Cham.

Generating Test Cases using Eclipse Environment: A Case Study of Mobile Application

Rosziati Ibrahim¹, Nurul Ain Aswini Abdul Jan², Sapiee Jamel³, Jahari Abdul Wahab⁴
Department of Software Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia^{1,2}
Department of Information Security, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia³
Engineering R&D Department, SENA Traffic Systems Sdn. Bhd, Kuala Lumpur, Malaysia⁴

Abstract—In Software Development Life Cycle (SDLC), there are four phases involved. They are analysis, design, implement and testing. Testing is done to ensure the functionalities of the system are correct. There are many approaches to software testing. It is usually divided into two approaches: manual testing or automatic testing. However, these days, with the rapidly advanced technology, performing software testing manually has become hugely laborious but still doable. Therefore, experts of the software development field are beginning to go for automatic testing. This paper presents a case study of mobile application and discusses how test cases can be generated automatically from the application using different automatic tools. Three software testing tools have been used to generate test cases automatically. The results from generating test cases automatically from these three tools are then being compared together with the results of generating test cases using manual testing technique.

Keywords—Software testing; automation testing; test cases; Eclipse environment

I. INTRODUCTION

In Software Development Life Cycle (SDLC), software testing is explained as the phase where a program is executed to be evaluated with the intention to find faults [1]. Although the SDLC is considered as an approach of efficient system development, software testing plays an important role as it assists in finding system deficiencies [2]. As such, testing is done to any software components, making it a vital process considering it aids in discovery of how good it works, validating the quality of the software system. To ensure that developed software components are in good quality, it is crucial to do software testing for the verification and validation to be done properly [3]. Considering how costly a software development project can amount to, testing becomes even more important, as prevention of even more highly cost of the software development. Therefore, it is important that the process is began at early stage during development [4] instead of being carried out by the end of the project development.

Software testing can be accomplished in two ways; either manually or automatically [5]. Manual testing is carried out by software testers without the help of any tools; it is a testing method which is most primitive compared to its peers [6]. On the other hand, contrary to manual testing, automatic testing is

performed with assistant from automated testing tool whereby test cases will be generated [7]. The performance capability and functionality of all test cases are to be justified. Testing tools are highly required to perform automatic testing. It plays a crucial role during the testing phase of the SDLC [8]. Several known tools include Robotium [9], Appium and Selenium [10].

This research study main aim is to generate test cases automatically from the existing tools and compared the time taken to generate test cases automatically among the tools. The case study is based on an existing Android mobile application called MyNetDiary [11]. The research shall be able to automate the process of generating test cases. There will be three tools used in the research which are JUnit4 [12], TestNG [13], and EPiT [14]. The results of time taken for each tool to generate test cases automatically will be compared together with the time taken to generate test cases using manual testing.

II. TECHNIQUES OF SOFTWARE TESTING

Generally, there are two ways to achieve software testing and those are by manual testing or automatic testing. The idea of manual testing is hugely primitive where the tests are executed in the absence of any tools [7]. Differing with manual testing is the automatic testing by which it is done with the help of automatic testing tools [14]. It is believed that by using automatic tools, the trend of automation testing has managed to have better usability, robustness, and correctness [8].

Software testing levels have been categorized into four levels [15]. These four levels include unit testing, integration testing, system testing, and acceptance testing which is shown in Fig. 1. Unit testing focuses on a software system's smallest element which is also known as modules; they are tested independently. Following after the unit testing is the integration testing where the main concept of it is testing the different integrated modules together. For most software project, the value of system testing being carried out is approximately up to 90 percent [16]. And then the last level of testing is the acceptance testing, performed by targeted end users [17]; it has variants of types which include alpha testing, beta testing, business acceptance testing, and the user acceptance testing.

This work was supported by SENA Project from MTUN under RMC, UTHM under Vote No. K234.

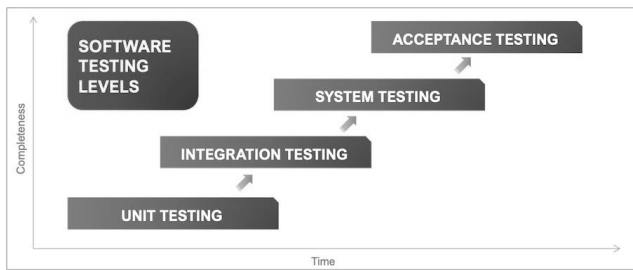


Fig. 1. Software Testing Level [15].

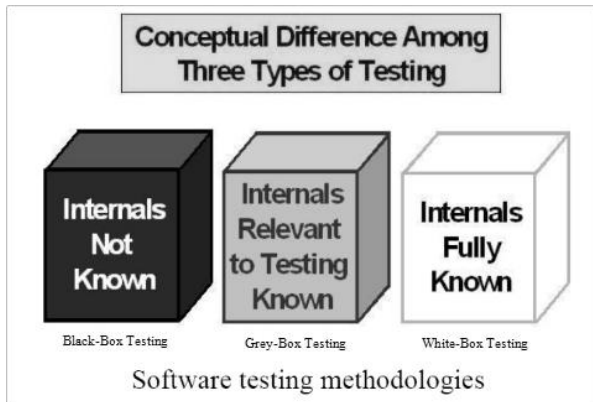


Fig. 2. Software Testing Methodology [17].

Fig. 2 shows the visualisation of three common software testing approaches [17]. Over the time period of rapid software development expansion, the common software testing techniques known to most are the black-box testing, white-box testing, and the grey-box testing methods. The grey-box method is a combination of the black-box and white-box testing methods [18].

A. Black-box Testing

Going by many other names such as behavioral testing and functional testing, black-box testing is usually driven without test models or even precise formal documented specifications [18]. The idea of black-box testing is the software testers do not know which of the system's component is being tested. As shown in Fig. 2, the idea of black-box testing is where the users or testers are without knowledge of the system's internals. The testing method concept is accepting inputs and producing expected outputs; black-box testing method borders on the foundation aspects of the system [19].

B. White-box Testing

There are many other nicked names to white-box testing method. Some of them include clear-box testing and glass-box testing. Just as the visual on Fig. 2 suggests, it is a testing method whereby the internals of the system are fully known [20]. As its nickname (clear-box testing), the back end of the system (or its components) is known to testers making it highly efficient in bugs-detection [21]. However, in large-scale software systems, this method is seldom used.

C. Grey-box Testing

Being the combination of black-box testing and white-box testing is the grey-box testing technique [22]. Fig. 2 shows the internals of the system is relevant to the testing being carried

out known by the testers. The concept of grey box is commonly known of testers having bits of internal working but going against its specifications [17]. The method typically applies reverse engineering but is not categorized as biased and intrusive; therefore the testers are not inclined to gain access on the internal source code.

III. RELATED WORK

Li *et al.* [23] present DroidBot, an automatic software testing tool which is compatible to most Android mobile apps. DroidBot is said as something that is lightweight and test on UI-guided input generators. It does not require any instrumentation. DroidBot also makes use of malware analysis as it uses a model-based generator that has information about app under test (AUT) from device at runtime, enabling it to trigger sensitive behaviours.

Alotaibi, & Qureshi [10] discuss a new framework to be used for automation testing on mobile application which will be using the Appium framework. According to them, in order to ensure high performance application within a short-given time, the automation of software testing is highly necessary. They specifically discuss Appium as it is considered as a power tool that helps in delivering features. What Appium does, to be precise, is the direct automation on mobile devices. It supposedly works for almost all of hybrid, native, applications of mobile-web for iOS and even Android.

Mao, Harman, & Jia [24] introduce Sapienz which is an Android testing approach that has significantly performed better than even the widely-used tool known as Android Monkey. According to them, Sapienz is better than Monkey is due to the fact that Monkey does automation testing in a deliberate unintelligent way of randomness. Sapienz, on the other hand, is a new automated testing that combines traditional automated testing with the quirks of expanding it to Android testing.

Dolan-Gravitt *et al.* [25] focus on PANDA's four principal criterion; the system's ability to record/replay, the system's plugin architecture, the system's capability in single analysis execution process on multiple architectures, and lastly the ability of Android systems emulation. PANDA is versatile and has simplicity, allowing support of new myriad of architectures and devices with no extra labour. The replay method itself is able to overcome the complexity of operating systems as it is able to record boot for myriads of operating systems. The system is more widely received considering its full repeatability features, a big convenience for dynamic analysis. Hence, considering PANDA is not focused solely on record and replay, it is adequately different than QEMU 2.1.0's numbers just as shown on the table below. However, PANDA takes almost the same amount of time as QEMU 2.1.0.

Hussain, Razak, & Mkpjojogu [26] discuss the perceived usability sentiments regarding the automated testing tools that exist for mobile testing. They discuss that many mobile application developers are using automated testing tools these days and that include MonkeyTalk, Robotium, and more. They state how it is no longer foreign that automated testing tools are gaining trend as it greatly reduces the time taken to

conduct the process of testing, excluding errors, and even omitting possible errors due to human factor. They argue how it has become highly important for automated testing tools to be of good usability as automated testing tools should not only support either native or hybrid, but they shall be able to do both. And that includes for Android and iOS.

Rosziati Ibrahim *et al.* [14] discuss the automatic testing tool called EPiT for generating test cases automatically. EPiT is a plug-in tool that can be installed in Eclipse environment. EPiT has a parser that reads the source codes line by line and then extracts all the attributes and functions from the classes and finally generates the test cases of all functions automatically.

Salihu *et al.* [27] propose a model to generate test cases from mobile application based on GUI. AMOGA framework is used for the generation of test cases with two important algorithms embedded within the framework. They are greedy algorithm and crawler algorithm.

IV. UML SPECIFICATION

UML diagrams are considered as the de-facto standard tool being used for the documentation of object-oriented modelling [28]. Two diagrams have been used for this project. They are use-case diagram and class diagram.

A. Use-case Diagram

Fig. 3 shows the use-case diagram of the research study. Based on Fig. 3, the actor is a user who can execute the tool in order to read the source code files of the case study. After doing so, it will be able to extract the classes and interface information, as well as checking the functions dependency. At last, it will generate the test cases.

B. Class Diagram

The class diagram portrays the classes that are going to be implemented during the development cycle. Fig. 4 shows the class diagram of this research study.

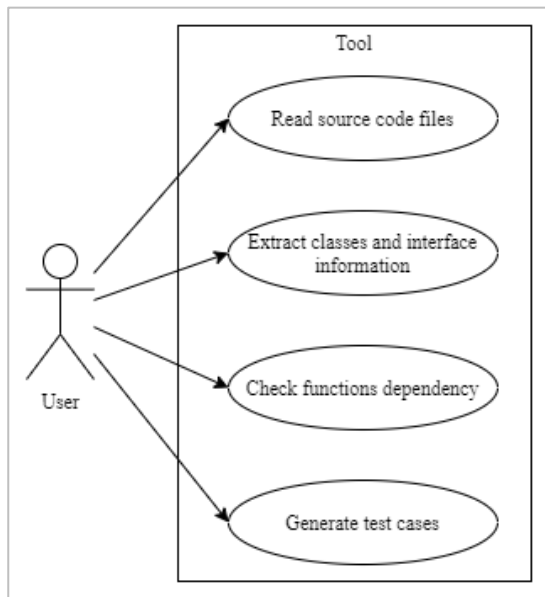


Fig. 3. Use-Case Diagram.

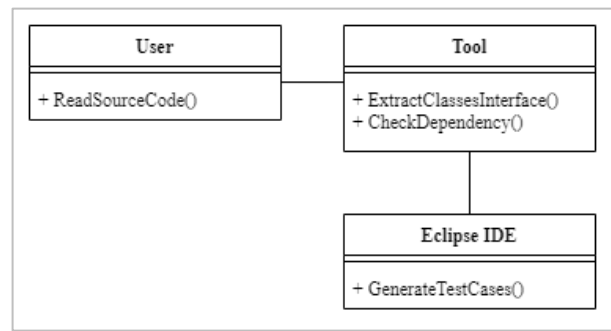


Fig. 4. Class Diagram.

Based on Fig. 4, it shows the specific of which methods belong either to the user, tool or the Eclipse IDE itself. The diagram does not exactly illustrate the directional work flow of the testing but it shows the classes that are being used for the implementation.

V. RESEARCH METHODOLOGY

The research study follows a specific process that consists of four stages to be carried out in order. For this research study, it will include a total of four major stages, the first stage being the requirement analysis. Next, it is followed by the design, implementation, and testing stages in an orderly manner. These four phases are illustrated in Fig. 5.

Based on Fig. 5, the requirements analysis stage is critical to this research study. As stated by Shukla, Pandey, & Shree [29], many other phases depend on requirements engineering and that includes the design, coding and testing. In this research study, this phase includes identifying the necessary tools and requirements needed. After identification, the requirements needed have been noted. The case study of the research is based on an Android mobile application which is MyNetDiary [11]. From MyNetDiary, the scope is further narrowed down to its 3 modules. The platform used for the research development is Eclipse IDE with the implementation of the Java programming language. Several other software and plugins are required for this research. As the codes of MyNetDiary mobile application cannot be fully obtained, it is determined that the software testing technique used is grey-box testing.

As the analysis phase, design phase is also included in SDLC. On a generic sense, during the design phase, the technical details of a software project are discussed, and this usually comprises of several aspects such as the technologies to be used, constraints, design approach, and so forth [30].

For the implementation phase, Fig. 6 shows the steps for implementing the tool.

Based on Fig. 6, the implementation process begins with first reading the source code file of the case study. Once the source codes are obtained, the automated testing tools which are running on Eclipse IDE will identify the classes and functions to be extracted. After that, the automated testing tools will begin generating the test cases automatically and the time taken for each of the tools and techniques will be observed, and recorded. For each software, testing methods, both manual testing and automation testing; the tests will be

run a total of 5 times for each Module 1, Module 2, and Module 3. This was done in order to get the optimal and most accurate data for the research. Lastly, the evaluation of time taken between the manual and automated testing will be made.

A. Manual Testing Flowchart

There are three basic activities to be done during the manual testing process as shown in Fig. 7(a). The case study file will first be run and executed, and then software tester will start inserting inputs. The time taken for the process to generate test cases will be recorded.

B. Automatic Testing Flowchart

Similar to the previous process of manual testing, automatic testing also follows several steps on generating test cases as shown in Fig. 7(b). The flowchart consists of four activities. The step begins with source code files of the case study being read. Its classes and interface information will be extracted, and the functions dependency will also be checked. Lastly, the test cases will be automatically generated by the selected tools.

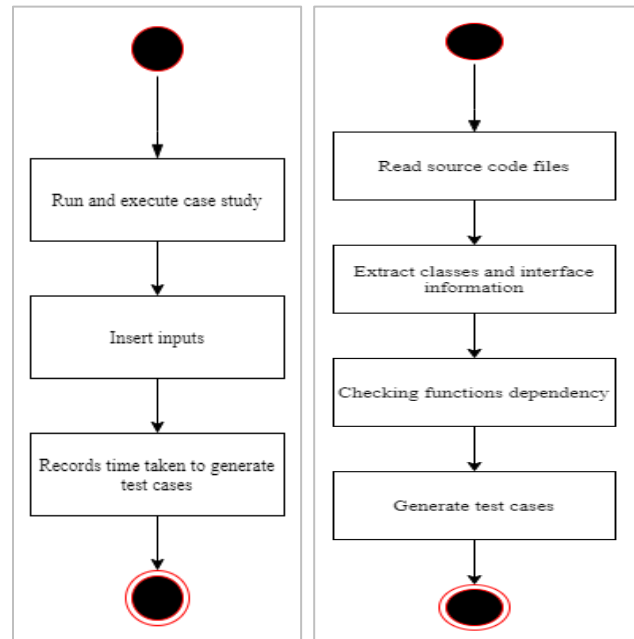


Fig. 7. (a) Steps for Manual Testing ; (b) Steps for Automatic Testing.

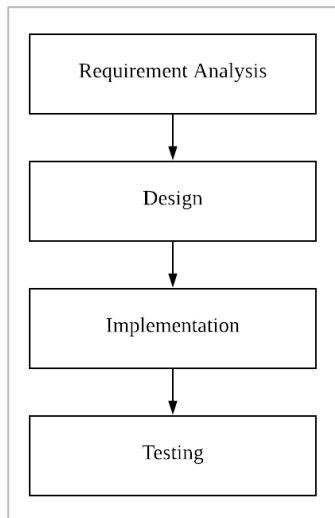


Fig. 5. Research Process.

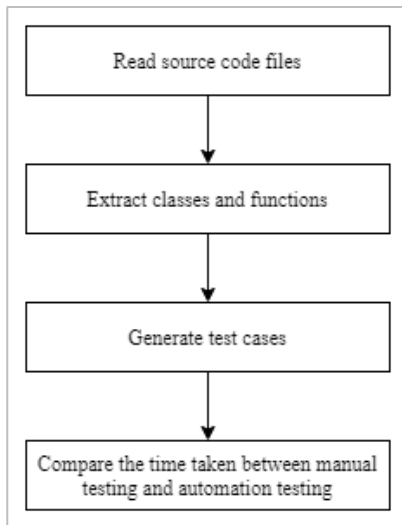


Fig. 6. Implementation Process.

VI. RESULTS AND DISCUSSION

Based on MyNetDiary [11], three modules have been used in order to generate the test cases. Table I shows the details of these three modules.

The data recorded from all the tests run during the research have been tabulated as each module is run at least five times for each respective automatic testing tools. The formula used to calculate the average time taken of tests run is:

$$\mu = \frac{\sum T_m}{5} \tag{1}$$

where \sum indicates the summation of the time taken to run for each module.

A. Manual Testing Results

Table II shows the calculation of data on the results of time taken to manually generate the test cases for all three modules.

Fig. 8 is the graphical diagram from Table II. It depicts the value of the average time taken to generate test cases manually for Module 1, Module 2, and Module 3. It took 21.884s, 13.672s, and 15.642s to generate the test cases for Module 1, Module2, and Module 3, respectively.

TABLE I. DETAILS MODULES FOR THE CASE STUDY

Module	Details
Module 1	Module of Calorie, BMI and Water
Module 2	Module to calculate the amount of calorie consumed from the different meals
Module 3	Module to calculate the amount of calorie burn ffrom different exercises

TABLE II. MANUAL TEST RUN ON THE CASE STUDY

Module	No. of Test	Time Taken (s)	Average Time Taken (s)
1	1	23.490	21.884
	2	22.080	
	3	20.920	
	4	21.710	
	5	21.220	
2	1	13.220	13.672
	2	13.720	
	3	13.600	
	4	14.240	
	5	13.580	
3	1	15.770	15.642
	2	15.600	
	3	15.590	
	4	15.570	
	5	15.680	

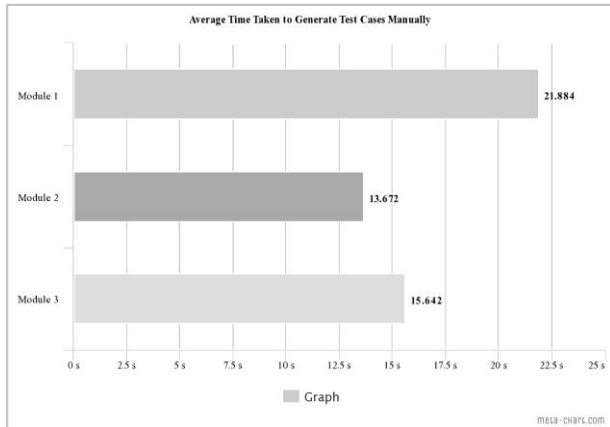


Fig. 8. Average Time Taken to Generate Test Cases Manually.

B. JUnit4 Testing Results

Table III shows the calculation of data on the results of time taken for Junit4 [12] to generate the test cases automatically for all the three modules.

Fig. 9 is the graphical diagram from Table III. It depicts the value of the average time taken to generate test cases automatically for Module 1, Module 2, and Module 3. It took 1.363s, 1.093s, and 0.598s to generate the test cases for Module 1, Module2, and Module 3, respectively.

C. TestNG Testing Results

Table IV shows the calculation of data on the results of time taken for TestNG [13] to generate the test cases automatically for all three modules.

Fig. 10 is the graphical diagram from Table IV. It depicts the value of the average time taken to generate test cases automatically for Module 1, Module 2, and Module 3. It took 0.016, 0.014s, and 0.020s to generate the test cases for Module 1, Module2, and Module 3, respectively.

TABLE III. TEST RUN ON CASE STUDY USING JUNIT4

Module	No. of Test	Time Taken (s)	Average Time Taken (s)
1	1	1.731	1.363
	2	1.361	
	3	1.191	
	4	1.460	
	5	1.070	
2	1	1.288	1.093
	2	1.099	
	3	1.054	
	4	1.080	
	5	0.943	
3	1	0.690	0.598
	2	0.553	
	3	0.578	
	4	0.625	
	5	0.544	

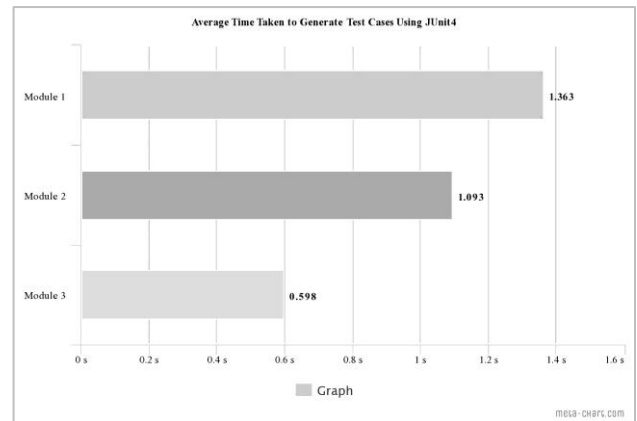


Fig. 9. Average Time Taken to Generate Test Cases Automatically using Junit4.

TABLE IV. TEST RUN ON CASE STUDY USING TESTNG

Module	No. of Test	Time Taken (s)	Average Time Taken (s)
1	1	0.013	0.016
	2	0.021	
	3	0.013	
	4	0.022	
	5	0.013	
2	1	0.014	0.014
	2	0.012	
	3	0.012	
	4	0.019	
	5	0.013	
3	1	0.020	0.020
	2	0.020	
	3	0.020	
	4	0.017	
	5	0.022	

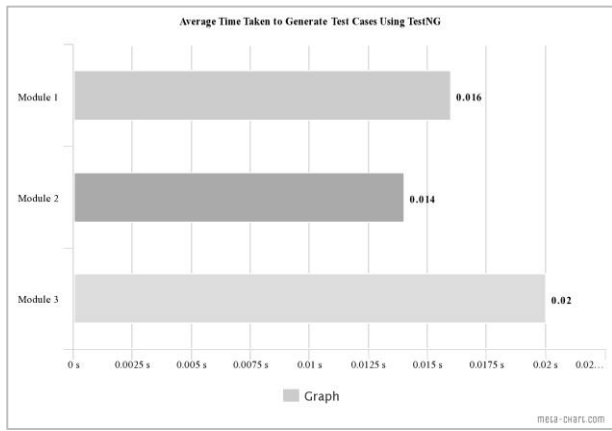


Fig. 10. Average Time Taken to Generate Test Cases Automatically using TestNG.

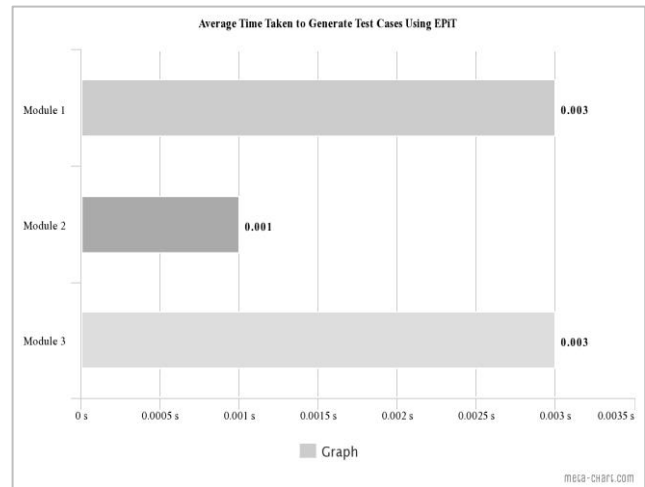


Fig. 11. Average Time Taken to Generate Test Cases Automatically using EPiT.

D. EPiT Testing Results

Table V shows the calculation of data on the results of time taken for EPiT [14] to generate the test cases for all the three modules.

Fig. 11 is the graphical diagram from Table V. It depicts the value of the average time taken to generate test cases automatically for Module 1, Module 2, and Module 3. It took 0.003s, 0.001s, and 0.003s to generate the test cases for Module 1, Module2, and Module 3 respectively.

Fig. 12 shows one of the runtime on Module 2 using EPiT. It took only 0.0001s to generate the test cases automatically from Module 2.

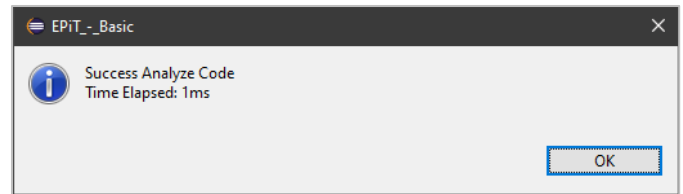


Fig. 12. EPiT Time Elapse for Module 2

E. Comparative Analysis

Table VI and Fig. 12 are the tabulated data and graphical representation of all testing methods. The time taken to generate the test cases using manual testing takes a significantly longer time than the time taken for the automatic testing tools to generate the test cases. This is clearly shown in Table VI.

TABLE VI. MANUAL VS AUTOMATIC TEST RUN ON CASE STUDY

Module	Manual Testing	JUnit4	TestNG	EPiT
1	21.884s	1.363s	0.016s	0.003s
2	13.672s	1.093s	0.014s	0.001s
3	15.642s	0.598s	0.020s	0.003s
	51.198s	3.054s	0.050s	0.007s

TABLE V. TEST RUN ON CASE STUDY USING EPiT

Module	No. of Test	Time Taken (s)	Average Time Taken (s)
1	1	0.005	0.003
	2	0.002	
	3	0.002	
	4	0.002	
	5	0.005	
2	1	0.001	0.001
	2	0.001	
	3	0.001	
	4	0.001	
	5	0.001	
3	1	0.007	0.003
	2	0.001	
	3	0.003	
	4	0.001	
	5	0.001	

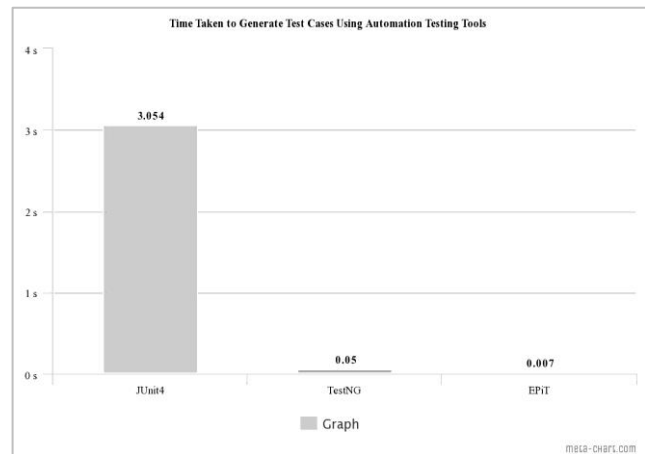


Fig. 13. Time Taken to Generate Test cases using Automatic Testing Tool.

From Table VI and Fig. 13, the time taken to generate test case using manual testing takes a significantly longer time than the time taken for the automatic testing tools to generate test cases. Among the three automation tools used, JUnit4 took the significantly greatest time which total reached more

than 3s. Meanwhile TestNG only took 0.05s to generate all test cases for all modules. Meanwhile, EPiT took the shortest time at only 0.007s.

Regarding the differences in time taken to generate test cases of the modules, this can be justified on the code lines of the case study. While the case study has simple time complexity of $O(1)$, the total number of lines for each module significantly differs with Module 1 having the most number of lines written, followed by Module 2, and Module 3. This causes for the time taken to generate test cases to differ from each of the respective modules. Beside from that, we can conclude that automatic testing is definitely better than manual testing. However, it needs to be noted that manual testing cannot be simply abandoned as it is still necessary for several tasks in any software development projects.

From Table VI, it is noted that manual testing has the biggest time difference compared to the others, which is just as expected. This is because manual testing demands a lot of resources which is one of them is the time resource [31]. Among the three automated testing tools, it is noted that the differences of time taken to generate test cases between JUnit4 and TestNG, as well as EPiT; JUnit4 takes the longest time. In one paper, Kumbhar, Gavekar, & Kulkarni [32] stated that JUnit is quite a lacking tool in generating test result compared to other testing tools. Meanwhile, it is no surprise that TestNG took shorter time than JUnit4 in generating the test cases, as according to Jacob and Karthikevan [33]. EPiT [14] is the latest software testing tool that has the shortest time to generate test cases automatically for the three modules. EPiT uses the algorithm in [34] in order to reduce the redundancy of test cases generated.

VII. CONCLUSION

This paper has discussed and compared the three automatic tools namely Junit4, TestNG and EPiT for generating test cases automatically. All three tools are plugged into Eclipse IDE. The time taken to generate the test cases has been compared among the tools. After the tests are run, it has been observed that JUnit4 took the longest time to generate all test cases, the time taken being almost up to 3s. Meanwhile TestNG only took 0.05s to generate all test cases for all modules. On the other hand, EPiT took the shortest time at only 0.007s. Therefore, EPiT gives the shortest time in order to generate test cases automatically. Beside from that, we can conclude that automatic testing is definitely better than manual testing. However, it needs to be noted that manual testing cannot be simply abandoned as it is still necessary for several tasks in any software development projects.

ACKNOWLEDGMENT

This project is funded by the Ministry of Higher Education Malaysia (MOHE) under the Malaysian Technical University Network (MTUN) grant scheme Vote K234 and SENA Traffic Systems Sdn. Bhd.

REFERENCES

- [1] Myers, G. J., Sandler, C., & Badgett, T. (1979). The art of software testing, JohnWiley & Sons. Inc, Canada.
- [2] Jindal, T. (2016). Importance of Testing in SDLC. International Journal of Engineering and Applied Computer Science (IJEACS), 1(02), 54-56.
- [3] Souza, É. F. D., Falbo, R. D. A., & Vijaykumar, N. L. (2017). ROoST: reference ontology on software testing. *Applied Ontology*, 12(1), 59-90.
- [4] Bertolino, A., & Marchetti, E. (2005). A brief essay on software testing. *Software Engineering, 3rd edn. Development process, 1*, 393-411.
- [5] Afrin, A., & Mohsin, K. (2017). Testing approach: Manual testing vs automation testing. *Global Sci-Tech*, 9(1), 55-60.
- [6] Patidar, R., Sharma, A., & Dave, R. (2017). Survey on Manual and Automation Testing strategies and Tools for a Software Application. *International journal of advanced research in computer science and software engineering*, 7(4), 10.
- [7] Anjum, H., Babar, M. I., Jehanzeb, M., Khan, M., Chaudhry, S., Sultana, S., ... & Bhatti, S. N. (2017). A comparative analysis of quality assurance of mobile applications using automated testing tools. *International Journal of Advanced Computer Science and Applications*, 8(7), 249-255.
- [8] Kochhar, P. S., Thung, F., Nagappan, N., Zimmermann, T., & Lo, D. (2015, April). Understanding the test automation culture of app developers. In *2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST)* (pp. 1-10). IEEE.
- [9] Zhu, Y., Hou, Y., & Wang, B. (2015). Application of automatic test tool Robotium for Android [J]. *Information Technology*, 10, 198-200.
- [10] Alotaibi, A. A., & Qureshi, R. J. (2017). Novel Framework for Automation Testing of Mobile Applications using Appium. *International Journal of Modern Education & Computer Science*, 9(2).
- [11] MyNetDiary.com. (2021) "Calorie Counter - MyNetDiary, Food Diary Tracker". Retrieved December 2020, from: <https://play.google.com/store/apps/details?id=com.fourtechnologies.mynetdiary.ad&hl=en&gl=US>
- [12] Junit4 (2021). Retrieved December 2020 from: <https://junit.org/junit4/>
- [13] TestNG (2020). Retrieved December 2020 from: <https://testng.org/doc/index.html>
- [14] Rosziati Ibrahim, Ammar Aminuddin Bani Amin, Sapiee Jamel, Jahari Abdul Wahab (2020). EPiT: A Software Testing Tool for Generation of Test Cases Automatically. *SSRG International Journal of Engineering Trends and Technology*, 2020, 68(7), 8-12. DOI:10.14445/22315381/IJETT-V68I7P202S
- [15] *Software Testing Levels*. (2020). Software Testing Fundamentals. Retrieved December 2020 from: <https://softwaretestingfundamentals.com/software-testing-levels/>
- [16] Jan, S. R., Shah, S. T. U., Johar, Z. U., Shah, Y., & Khan, F. (2016). An innovative approach to investigate various software testing techniques and strategies. *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)*, Print ISSN, 2395-1990.
- [17] Kassab, M., DeFranco, J. F., & Laplante, P. A. (2017). Software testing: The state of the practice. *IEEE Software*, 34(5), 46-52.
- [18] Jamil, M. A., Arif, M., Abubakar, N. S. A., & Ahmad, A. (2016, November). Software testing techniques: A literature review. In *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M)* (pp. 177-182). IEEE.
- [19] Lawanna, A. (2014). The theory of software testing. *AU Journal of Technology*, 16(1), 35-40.
- [20] Sneha, K., & Malle, G. M. (2017, August). Research on software testing techniques and software automation testing tools. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 77-81). IEEE.
- [21] Mailewa, A., Herath, J., & Herath, S. (2015, April). A Survey of Effective and Efficient Software Testing. In *The Midwest Instruction and Computing Symposium*. Retrieved from http://www.micsymposium.org/mics2015/ProceedingsMICS_2015/Mailewa_2D1_41.pdf.
- [22] Poullova, P., & Klimova, B. (2018). Automated Software Testing—A Case Study. *Advanced Science Letters*, 24(4), 2578-2581.
- [23] Li, Y., Yang, Z., Guo, Y., & Chen, X. (2017, May). Droidbot: a lightweight ui-guided test input generator for android. In *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)* (pp. 23-26). IEEE.

- [24] Mao, K., Harman, M., & Jia, Y. (2016, July). Sapienz: Multi-objective automated testing for Android applications. In *Proceedings of the 25th International Symposium on Software Testing and Analysis* (pp. 94-105).
- [25] Dolan-Gavitt, B., Hodosh, J., Hulin, P., Leek, T., & Whelan, R. (2015, December). Repeatable reverse engineering with PANDA. In *Proceedings of the 5th Program Protection and Reverse Engineering Workshop* (pp. 1-11).
- [26] Hussain, A., Razak, H. A., & Mkpojiogu, E. O. (2017). The perceived usability of automated testing tools for mobile applications. *Journal of Engineering, Science and Technology (JESTEC)*, 12(4), 89-97.
- [27] Salihu, I.A., Ibrahim, R., Ahmed, B.S., Zamli, K.Z. Usman, A. (2019). "AMOGA: A Static-Dynamic Model Generation Strategy for Mobile Apps Testing". IEEE Access. 2019. DOI:10.1109/ACCESS.2019.2895504.
- [28] Pender, T. (2003). *UML 2 Bible*. John Wiley & Sons.
- [29] Shukla, V., Pandey, D., & Shree, R. (2015). Requirements Engineering: A Survey. *Requirements Engineering*, 3(5), 28-31.
- [30] Rani, U., Barjtya, S., & Sharma, A. (2017). A detailed study of Software Development Life Cycle (SDLC) models. *International Journal Of Engineering And Computer Science*, 6(7).
- [31] Garousi, V., & Mäntylä, M. V. (2016). When and what to automate in software testing? A multi-vocal literature review. *Information and Software Technology*, 76, 92-117.
- [32] Kumbhar, M., Gavekar, V., Kulkarni, A. (2020). Performance Testing Tools: A Comparative Study of QTP, Load Runner, Win Runner and JUnit.
- [33] Jacob, A., & Karthikeyan, A. (2018, March). Scrutiny on Various Approaches of Software Performance Testing Tools. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 509-515). IEEE.
- [34] Ibrahim, R., Ahmed, M., Nayak, R., Jamel, S. (2020). "Reducing redundancy of test cases generation using code smell detection and refactoring". *Journal of King Saud University – Computer and Information Science*, Volume 32, Issue 3, March 2020, pp 367-374. DOI:10.1016/j.jksuci.2018.06.005.

Traffic Accidents Detection using Geographic Information Systems (GIS)

Spatial Correlation of Traffic Accidents in the City of Amman, Jordan

Wesam Alkhadour¹, Jamal Zraqou², Adnan Al-Helali³, Sajeda Al-Ghananeem⁴

Departments of Civil Engineering, Computer Science, and Software Engineering, Isra University, Amman, Jordan

Abstract—The mission of reducing the number and severity of traffic accidents becomes the dominant target of road traffic safety management worldwide. The main objective of this work is to analyze traffic accidents in temporal and spatial frameworks in the capital city Amman and identify hotspot zones in the study area. Several statistical analyses are conducted using SQL to give insight into the temporal distribution of accidents and to identify the most revealing accidents based on several attributes such as the year of accidents, the severity of accidents, road type, and lighting environment which enables the authors to do further investigations on the more frequent accidents. GIS-based statistical and spatial analysis tools are utilized to examine the spatial pattern of accident distribution in the study area for three successive years, hotspots are identified for clusters of high concentrations. The Nearest Neighbor Index (NNI) is used to analyze the pattern of traffic accident distribution based on selective parameters. This was followed by identifying hotspot zones for regions that showed clustering using the optimal hotspot analysis tool. Experimental results showed clustering for all tested groups, and thus hotspots were detected for these accidents in the study area. The importance of this work is in providing a spatial understanding of accident distribution in the capital city of Amman which can help policymakers of road safety setting out efficient strategies for traffic safety management and find optimal solutions as required for factors causing such accidents.

Keywords—Geographic Information System (GIS); statistical tools; hotspots; spatial analysis; temporal analysis; road safety; traffic accidents; spatial correlation

I. INTRODUCTION

Traffic accident problem is major public health anxiety in the globe. Statistics showed that there are about 1.3 million deaths and 50 million injuries caused by road accidents each year [1].

The traffic authorities in most countries tend to control the traffic to reduce traffic accidents using several techniques such as: optimize traffic-light management, improve cycling infrastructure, enforce existing road traffic laws, improve perceptions of buses, extend residents' parking zones, use CCTV to monitor road conditions, charge for workplace parking, improve bus services, etc.

The impact of road accidents on both economy and society is massive. It is estimated that the cost of these accidents is about 518 billion USD worldwide [1, 2]. In developing countries such as the Kingdom of Jordan, people are more vulnerable to road accidents according to the global status report issued by the World Health Organization(WHO) on road safety

[3], Jordan has appeared in the set of the worst countries regarding road safety. The report revealed that the rate of deaths caused by road accidents is about 24.4 for each 100.000 of population.

Jordan's population experienced growth in the last decade to reach roughly 9.9 million [4] leading to growth in road accidents. Road crashes are considered the second cause of death [5] in Jordan. Most of the casualties of these accidents are passengers and pedestrians who represented (62.3%) of fatalities resulted from road accidents safety [3]. The region of interest of this study is shown in **Error! Reference source not found.**

Road and traffic safety planning is a life basic requirement in all societies. Safety information is essentials for addressing road and traffic safety needs. Informed decision-making of road safety issues is based on the registered crash information. The accident's information was collected from the police traffic department. The collected data then is used to analyze and identify the hazardous regions on road, then the pattern of accidents can be identified by applying engineering studies. Results of the analysis are used to implement the required improvements and road planning.

Geographic information systems tools are utilized for advancing safety planning [6]. Applications of GIS in the management of accident analysis and traffic safety were introduced in 1990. Several GIS-based accident information systems are devoted to managing and tackling traffic flow. These GIS-aided systems aim to decrease the number of accidents by identifying the hazardous locations which can be achieved by using the spatial data analysis and statistical analysis methods provided in GIS software Booth [7]. GIS software includes a variety of analysis tools such as density, proximity, cluster, pattern, spatial query analysis in addition to the ability to build customized models using the model builder technique.

GIS-based accidents information system overcomes tabular database information system in enabling a user to identify relationships which is very difficult to be achieved using tabular database systems this makes the GIS-based information system one of the powerful spatial information systems. The main components of a spatial information system are a tabular database representing accident data, a GIS system that connects the database to the corresponding locations, and a set of tools to analyze the attribute and spatial data. Results of analysis aids in road and traffic safety planning enhancements.

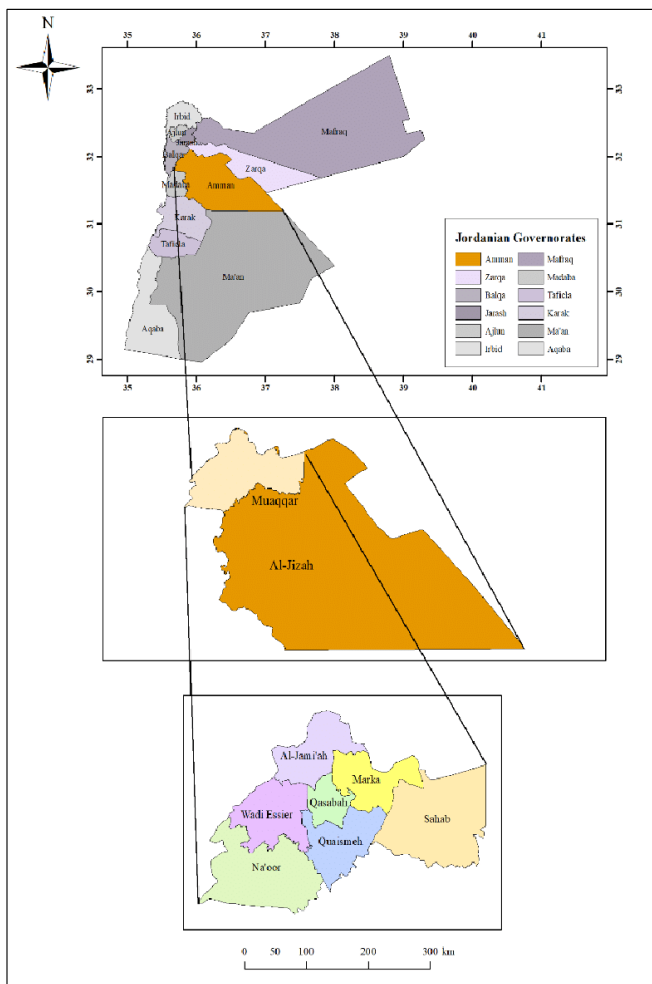


Fig. 1. The Studied Area of Amman.

Identifying accidents hotspots is a major part of most safety analysis studies, where locations with the highest proportions of total accidents are identified. There are two main categories for identifying hotspots, observed accidents rate, and expected accident rates. However, hotspots can be identified efficiently by using methods that incorporating both categories because accidents are considered random events and they can change in time and locations [11].

Several methods have been introduced for identifying hot zones [8-10]. Each study has advantages and drawbacks and is designed to address hotspots identification issues in a specific environment. The most widely used approaches for hotspot identification [11] is the kernel density estimation method (KDE) [12, 13], nearest neighbor hierarchical method (NNH) [14], and local indicator for spatial association (LISA) [15], where these methods indicate the spatial pattern of events, whether it is clustered, random or dispersed. Having the spatial pattern identified, the hotspots are then identified using various tools available in GIS software or any other software designed for this task.

The quality of results obtained from accident analysis is affected by several factors. One important factor contributing to the success of the spatial and statistical analysis is the accuracy

and reliability of spatial databases which are extracted from accident reports. A study was conducted in [16] on accident report forms used in many countries around the world, the study revealed that there are 99 different types of information according to the environment of the accidents.

Only a few studies have examined traffic accidents in a spatial framework in the city of Amman, Jordan. This study is aimed to examine accident data in both temporal and spatial frameworks based on several categories and consequently unravel the main zones of accident concentrations for categories that showed a high proportion of total accidents. The objectives of this study are addressed by first conducting temporal statistical analysis for accident data using SQL based on several categories such as year of accident, severity, road type, and lighting condition, this highlights the recurrence of accidents related to some categories other than accidents related to other categories. GIS-based tools are then utilized to find out the spatial pattern of these accidents, then identify the hotspot zones for patterns showing clustering. This can help in providing more understanding of the geography of traffic accidents and the spatial correlation between these accidents in the study area and as a result, help interested parties in setting precautions and optimal solutions to safety problems.

The remaining sections are organized as follows: Section 2 provides an extended literature review of recent research on the hotspot analysis and the main causes of the accidents. Section 3 describes the dataset that is going to be used to conduct the experiments and evaluations. Section 4 presents the authors' approach to analyze accidents data based on various attributes, examining the distribution of accidents in the study area and identifying hotspots, and finally concluding this work and suggesting the future works are presented in section 5.

II. LITERATURE REVIEW

In general, this section covers the literature review from different sources related to the studies of traffic and road safety that are focused on the factors affecting safety. Several factors contribute to road and safety deficiency some of these factors are caused because of the structure of the road, geographical and weather conditions and deficiency in the lighting system of roads [17]. Spatial databases are manipulated using GIS technologies which result in identifying the relationships between these spatial phenomena.

Another study for accident analysis was introduced in [7], the study objective was to build a GIS-based system for identifying hotspots in Afyonkarahisar in Turkey and factors causing these accidents in these areas with statistical analysis methods so suitable solutions can be applied by road safety specialists. Also, a Safety Evaluation Method for Local Area Traffic Management (SELATM) was introduced in [7]. Their system is mainly based on GIS technology for analyzing accidents pattern.

Another traffic accident analysis system was introduced in [18]. The database of accidents, structures of roads, and facilities of road accessories can be managed and analyzed using the developed system. As a result, rates of accidents along with their frequencies can be identified. The work in [19] introduced a GIS-based system for accident analysis. The

system is designed for identifying the accident location besides that the rank of the accident can be identified as well. The system enables the user to input and retrieves a database related to accidents and manipulate statistical analysis on a specific accident location.

An aided GIS model for identifying hotspot location of road accidents was developed in [20]. The data used for analysis and determining the locations of the hotspots are the XY coordinates of road accidents obtained from traffic police. The developed model along with the other three models for scheduling and operating system of the enforcement camera, controlling and balancing the traffic load in the courts and a model for analyzing videos contribute to reducing the number of accidents and fatality.

A GIS-based Analysis to Identify the Spatiotemporal Patterns of Road Accidents in Sri Racha, Chon Buri, Thailand was developed in [21]. The analysis is performed by applying kernel density estimation (KDE), and Ripley's K-function tools in ARC-GIS to determine the distribution and pattern of the accidents. The database used in this study is the accident data from 2017 to 2020. Several scales for clustering the spatiotemporal pattern of accidents were used. Clustering the spatial distribution of different accident types was performed at different distances. Experimental results show that there are three main areas in the studied area with high-density accidents.

An innovative spatial-auto correlation-based method for identifying road accident hot zones was presented in [22]. A new method based on ARC-GIS software and spatial autocorrelation algorithm was built for identifying hotspots of accidents taking into consideration both the properties and attributes of the accident. The proposed method is applied on-road sections divided into several 100 meters long, as a result, the location of where the accident is identified regardless of the rate of the accident in this location. However, further work can be done for classifying the hot zones based on different rules such as the severity of an accident.

A study for identifying hot zones of accidents and analyze the relationship between these accidents and land use was introduced in [11]. The first step of examining the hot zone was performed using ARC-GIS tools. The analysis of hot zone is mainly applied for three categories, severity category which includes two groups, causes of crash occurrences which includes seven dominant causes, and three major types of accidents frequently happening in the study area of Dammam city, Kingdom of Saudi Arabia (KSA). The next step of identifying the relationship was achieved by applying a GIS-based geographically weighted regression (GWR) method to identify the relation of accident and density of population and type of land use in the crash area. Although this study contributes to analyzing the complex relationship between the crash and land use in the crash neighborhood, various detailed features of the neighborhood environment could be considered to advance the existing analysis tools.

A GIS-based system is incorporated into fuzzy logic to predict the hot zones of accidents was proposed in [23]. The spatial-temporal analysis was applied to examine fatality and injury groups in the context of accident severity. The prediction process is performed by applying the Fuzzy Overlay Method

(FOM) and the Weighted Overlay Method (WOM). The results obtained from the previous step are verified using the density point tool. The proposed algorithm is evaluated using the database between 2013 and 2015 of Irbid city Jordan. Results show that there are 8 hot zones, five are main road intersections and three road sections were investigated to identify factors causing these accidents.

A GIS-based system incorporated with the Firefly Clustering algorithm was presented in [24] to identify hot zones of accidents. spatial analysis tools existing in ARC-GIS were used to find distances between accident points, while characteristics of accidents were identified by applying a Firefly Clustering algorithm. The performance of the developed method was evaluated by conducting a comparison between distances calculated using the GIS origin-destination (OD) cost tool and the Euclidean distance tool. Results show that number of hot spots is overestimated using Euclidean distance, particularly at intersection zones.

A GPS and Arc-GIS incorporated system was proposed in [25] to identify black zones in the city of Pristina, Kosovo. In this study, maps were created using GPS technology and compared to traditional methods of collecting accident data from location. Results of evaluating the proposed methods reveal that GPS technology in collecting accident data gives more accurate black spot identification rather than conventional methods of collecting record data of accidents.

A spatial accident information system was introduced in [26]. The system manipulates attributes characteristics of accidents obtained from accident reports using a set of spatial analysis tools provided in GIS. Hotspots of crashes are identified by applying stat crime programs based on the nearest neighbor hierarchical clustering technique [27]. Accidents along roadways were analyzed, interpolation of crashes is applied using the Crime stat program, which shows the presence of hot spots of accidents with high risk. The advantages of this system over other systems that it can be used for usual types of accident analysis and difficult types of analysis that cannot be performed with tabular system such as spatial selection. However, more accuracy in spatial information is required to improve the performance of the system.

A study for identifying crash hotspots based on GIS methods in Muscat city the capital of Oman was introduced in [28]. Various methods were employed to identify the crash hot spots including Kernel Density Estimation (KDE), Network-based K-Function, Network-based Nearest Neighbor Distance (Net-NND), spatiotemporal Hot-zone analysis, and Random Forest Algorithm (RF) and. Results confirm that road intersections influence road accidents more than other geometric features of the road. A cell-grid model was proposed in [29] for bicyclist risk maps in Manhattan, New York City, the Bayesian framework was utilized to develop a random parameter model which is used to correlate the cost of bicycle accident with land use, transportation and sociodemographic data. Findings confirm that the proposed method is superior to the Tobit model.

A prediction for heavy vehicle accidents hotspots clusters was introduced in [30]. The clustering is based on three criteria, and achieved using Moran's I spatial autocorrelation. The risk

along the network was estimated using Getis–Ord G_i^* statistic, findings reveal the existence of 22 segments remarked with heavy vehicle risk.

This paper presents a statistical analysis of traffic accidents in the city of Amman, Jordan for three successive years 2017, 2018, and 2019 based on different attributes such as severity, type of location where accidents have occurred, lighting environment of accidents' locations. The distribution of accidents is examined using GIS spatial tools. The hotspots are also identified using GIS tools for traffic accidents in each studied year and traffic accidents causing fatalities and serious injuries in each year as well. Moreover, the hotspots are identified for roads that registered the highest proportion of accidents. A new contribution presented in this paper is about adding an extra attribute on the collected data to show the status of the road if it sets on a transport road or not. Hence the hotspots can be analyzed by correlating road status parameters in terms of a transport road or not with other parameters provided with the data and described in the following section.

III. DATA

The accident data of the years 2017, 2018, and 2019 were obtained from the traffic institution in Amman the capital city of Jordan, these data were converted into an SQL server format. This could provide the ability to build relational algebra expressions to represent intended methods.

The attributes of the data are divided into 12 categories as follows:

- The date of the accidents.
- The location (latitude, longitude) of the accidents.
- The number of accidents for each location of accidents is going to be counted.
- Minor injuries.
- Intermediate injuries.
- Major injuries.
- The direction of lanes (one way, or bidirectional ways).
- Driving license type.
- Lighted road (yes or no).
- Types of mistakes that cause accidents.
- Type of vehicle (Passengers or Shipping).
- Type of accidents (Vehicle crash or Pedestrian crash).

The roads of interest can be selected based on the map of transport in the city of Amman as shown in **Error! Reference source not found.**

Also, the data collected in [31] described in **Error! Reference source not found.** were considered and being updated from the main source of Jordan Traffic Institute and Traffic Department.



Fig. 2. The Transport Map of a Region in Amman.

TABLE I. HAZARDOUS STREETS IN AMMAN [28]

#	Street Name	#Accidents	#Injuries	#Fatalities
1	Abu Alanda st	52	72	2
2	Kraibt AlSog st	24	28	2
..

IV. RESULTS AND DISCUSSION

A. Accidents Temporal Distributions

This section presents accident temporal variations in the studied region based on several categories: years, accident location, severity, and lighting environment.

Error! Reference source not found. shows the number of injuries, and deaths for the years 2017, 2018, and 2019. A total of 628,006 crashes were recorded, resulting in 33,439 injures and 897 deaths. Also, this figure shows that the year 2019 includes the highest number of accidents, injuries, and deaths. Consequently, an analysis of the data was required to justify the main causes and providing recommendations to take further actions by the authorities of road safety.

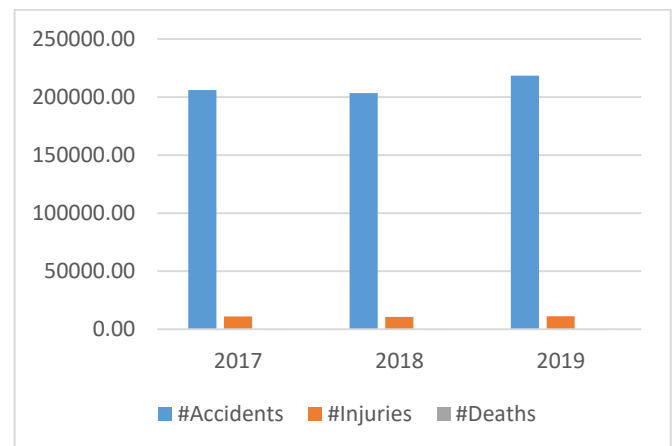


Fig. 3. Number of Accidents, Injuries, and Deaths in the Years of 2017, 2018 and 2019.

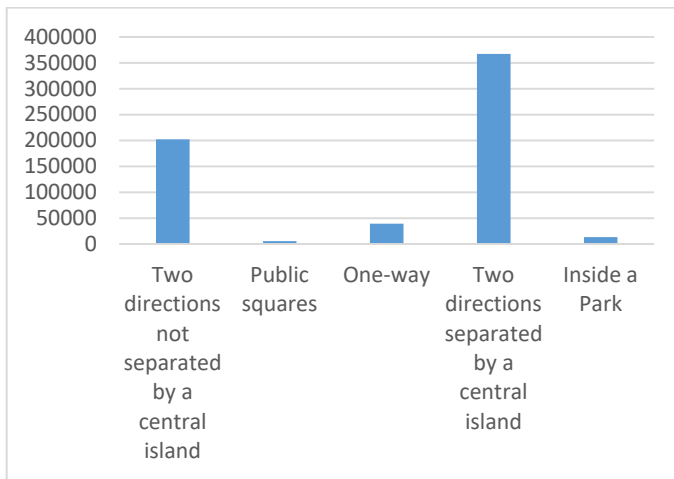


Fig. 4. Number of Accidents based on the Pathway of the Road.

Error! Reference source not found. shows the numbers of accidents according to the location of accidents. In this category, five major groups according to accident locations are taken into accounts (a road with two lanes separated by a central island, a road with two lanes not separated by a central island, one-lane road, public square, and inside a park). The accidents that occurred on roads with two lanes separated by a central island constitute the highest proportion of the total number of accidents which exceeded 350000 accidents. This is followed by 200000 accidents recorded on roads with two lanes not separated by an island. The three remaining groups represent a small proportion of the total number of accidents.

Another statistic for the category of accident locations along with the year category was performed, where the number of accidents according to accident location is calculated for each year separately. Results are shown in **Error! Reference source not found.** It can be observed that there is a serious problem with roads of two lanes, especially where lanes are separated by a central island, which is more vulnerable to traffic accidents. This high number of accidents stirred further investigations to find out factors contributing to accidents and taking preventive measures to tackle this problem.

Error! Reference source not found. shows the results of a statistical analysis based on severity category. In this category, four groups are considered (simple injuries, intermediate injuries, serious injuries, and fatality). There is almost a convergence between the total number of casualties in each studied period. However, 2017 year witnessed the smallest number of intermediate injuries while 2019 witnessed the largest.

Error! Reference source not found. shows statistical results based on the lighting category. In this category, several groups are studied including various parts of the day and lighting environment (day, rise, sunset, darkness, etc.). Results

show that most accidents have occurred during the daytime in the three studied years. The second-largest proportion of accidents occurred at night through adequate lighting conditions. A small proportion of accidents occurred during the night with not enough lighting and during sunset time. These results are consistent with the system of life in the studied area Amman, where most roads witness a heavy traffic volume during the day due to the high density of population in the capital city of the kingdom.

B. Identifying Hot Spots of Accidents Analysis

Although the accidents have been thoroughly examined based on various attributes using SQL, a spatial framework enables in identifying the pattern of accident distribution which provides a better understanding of road safety. Based on the type of spatial pattern of accidents, hotspots can then be identified. The hotspots can be defined as a concentration of accidents at or near a specific location. Analysis can be applied to investigate clusters of events that occur near each other. Hotspots then can be identified by using spatial analysis tools provided in GIS software, where several crashes occurring at a given location are counted. . An alternative way to identifying hot spots of crashes is using stat crime programs based on nearest neighbor hierarchical clustering [32].

In this work, the nearest neighbor index (NNI) is used to analyze the pattern of traffic accident distribution. The nearest neighbor index was calculated using the average nearest neighbor tool provided by ArcGIS v.10.8. The nearest neighbor is considered one of the popular methods used to identify point pattern and it is calculated using equation (1)

$$NNI = \frac{\bar{D}_{nnd}}{D_{ran}} \quad (1)$$

Where

D_{nnd} represents the average distance between each accident location and the nearest accident location.

$$D_{nnd} = \frac{\sum_{i=1}^n d_i}{n} \quad (1)$$

D_{ran} represents the expected distance based on the hypothesis that the accidents are distributed randomly in the studied area.

$$\bar{D}_{ran} = 0.5 \sqrt{\frac{A}{n}} \quad (3)$$

In equations (2) and (3), n represents the number of accidents in the studied area and A is the area of the studied region.

In this work, the spatial pattern of traffic accidents of the years 2017, 2018, and 2019 respectively are analyzed. Results show that the accidents are clustered for each studied year, where NNI is less than one for each year.

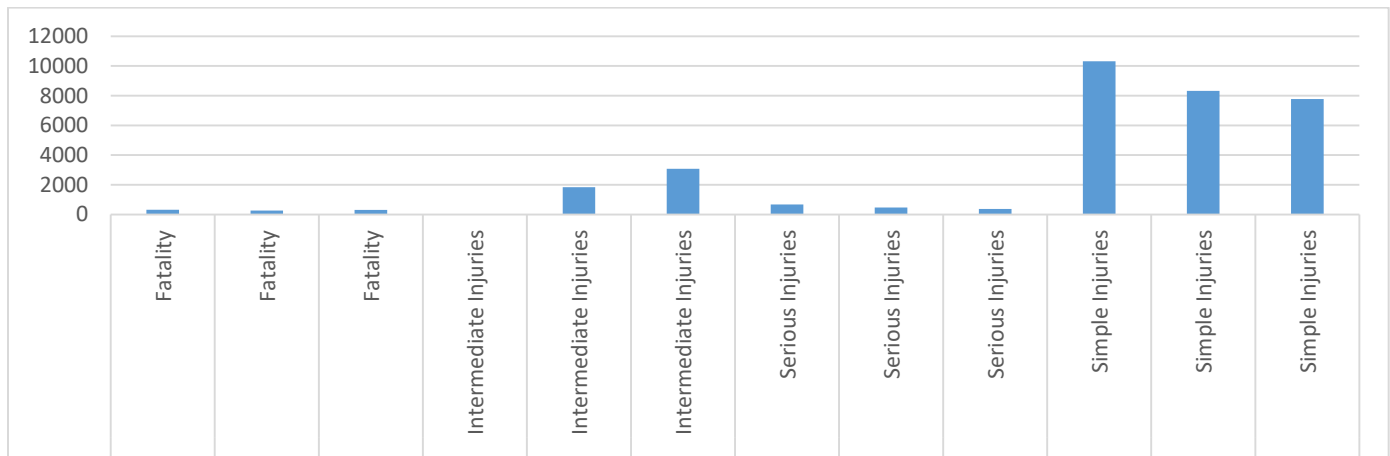


Fig. 5. Number of Casualties based on the Level of Severities for the Years of 2017, 2018 and 2019.

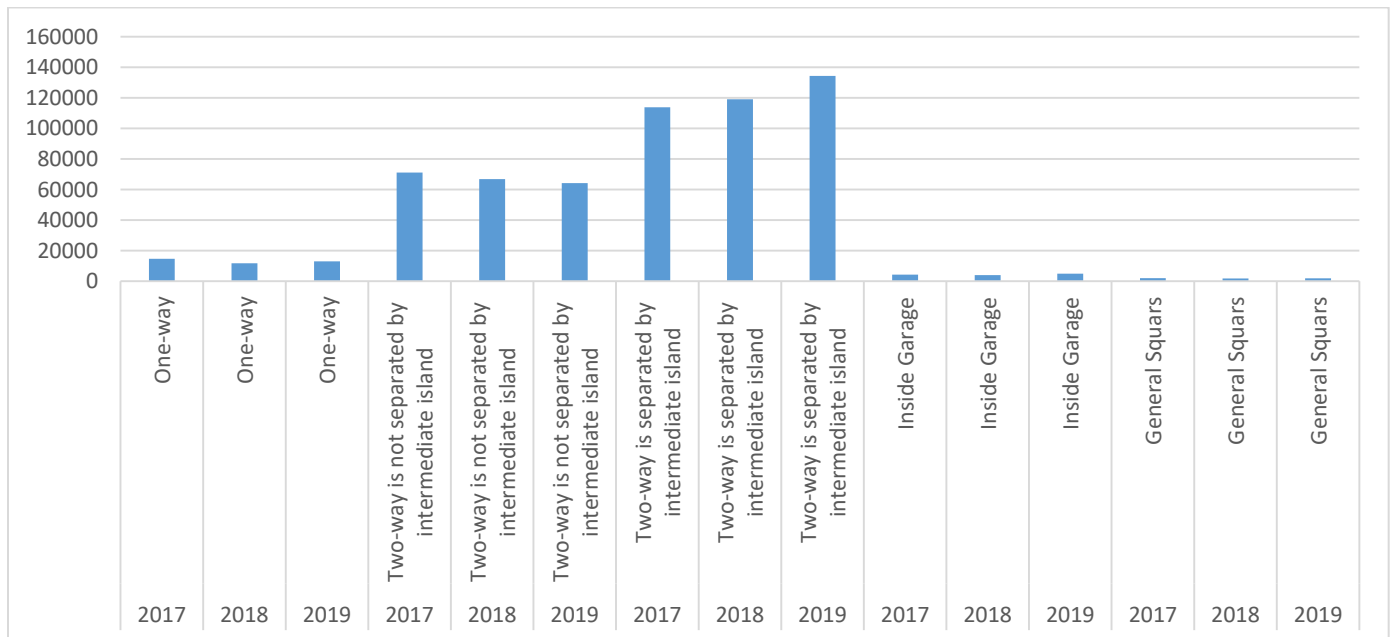


Fig. 6. Number of Accidents based on the Pathway of the Roads Classified by the Years of 2017, 2018 and 2019.

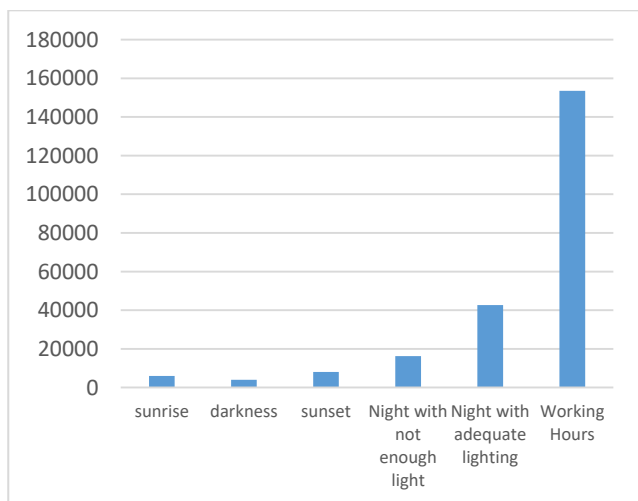


Fig. 7. Number of Accidents based on the Level of Street Lights.

Since the priority of this work is to mitigate the fatalities and casualties of road accidents, the nearest neighbor analysis is conducted for two severity groups (fatalities and serious injuries group) for each studied year. Results of the analysis showed that accidents causing fatalities are clustered as well, where NNI is less than one for all studied years as shown in **Error! Reference source not found.II**.

Regarding the serious injuries group, also (NNI) index indicates the clustering of accidents. However, the clustered accidents are not significant for the years 2017 and 2019, while there is a clustering of accidents resulting in serious injuries in the year 2018 with a small significance level.

As shown in the previous section the majority of traffic accidents were occurring on roads with two lanes separated by a central island. This is followed by accidents occurring on roads with two lanes that are not separated by an island. These results require further spatial analysis to find out the locations of these roads. The nearest neighbor analysis is performed on accidents

occurring in these two types of roads. Results show that the accidents are clustered in the studied area. Results are summarized in **Error! Reference source not found.II Error! Reference source not found.**

Since the results of all the examined accidents showed clustering status, hotspots are identified for accidents each year, for two severity groups, two types of roads. Hotspot zones are identified using the optimized hotspot analysis tool provided in ArcGIS. This technique is based on creating a fishnet polygon Cells. Each cell has a side length of 250 m. Results show that the hotspot areas were mapped with more than a 90% confidence level.

Hotspots maps were produced for the whole accidents occurring in each studied year 2017,2018 and 2019 as shown in Figure 8[a, b and c]. Findings clearly pinpoints the concentration of accidents in the same zones for each studied year which are mainly in and around the central regions of the study area. Other hotspots maps were produced for accidents causing fatalities for each studied year as shown in Figure 9[a,b, and c]. Results are similar to the aforementioned results in the manner of distribution of regions of accidents concentration, where hot zones are mainly distributed in and around the center

of the study area. However, accidents distribution that causing fatalities in year 2018 is noticeably less than accidents concentrations in year 2017 and 2019.

Regarding the accident resulting in serious injuries group, Findings reveals that there are no significant presence of hot spots in year 2017 and 2019 with small presence of accident hotspots in year 2018 as shown in Figure 10. Since accidents located on roads of two lanes show high significance clustering, hotspots maps were produced for this category of accidents as well. As shown in Figure 11[a, b and c], and Figure 12[a, b and c], hotspot zones distribution for both groups of accidents in this category are similar to each other , to the distribution of accidents related to severity category and to the general distribution of hotspot zones along the studied years.

The spatial analysis highlight evidence of spatial clustering and recurrence of traffic accidents in the central regions of the studied network. The findings confirm that roads directed to the center of the studied network are more vulnerable to traffic accidents other than roads that are located away from the center of the study area. This confirms that geography of the road and neighborhood elevate the risk of accidents concurrence.

TABLE II. NEAREST NEIGHBOR INDEX (NNI) VALUES AND THE NUMBER OF ACCIDENTS EACH YEAR

year	variable	#of accidents	Observed mean distance (m)	Mean/expected Random distance (m)	NNI	Z-value	# of accidents within the hot spots zones
2017	none	206050	7.06	288.12	0.02	-238.86	1049
2018	none	203511	1.21	91.38	0.01	-851.56	1109
2019	none	218445	1.58	98.18	0.02	-879.74	962
Severities							
2017	Fatalities	285	466.28	1771.80	0.26	-23.79	111
2018		238	423.09	1839.36	0.23	-22.72	84
2019		246	601.53	2581.29	0.23	-23.01	88
2017	Serious injuries	513	234.41	1155.75	0.20	-34.54	198
2018		405	320.84	1246.89	0.26	-28.59	153
2019		338	378.49	1977.76	0.19	- 28.44	121
Road Lanes							
2017	Two lanes separated by a central island	113825	0.79	88.25	0.009	-639.59	827
2018		119063	1.08	104.79	0.01	-653.28	851
2019		134405	1.32	121.92	0.01	-693.78	688
2017	Two lanes not separated by a central island	71180	2.76	146.68	0.02	-500.79	888
2018		66833	2.77	145.36	0.02	-485.14	984
2019		134405	3.35	170.20	0.027	- 475.37	819

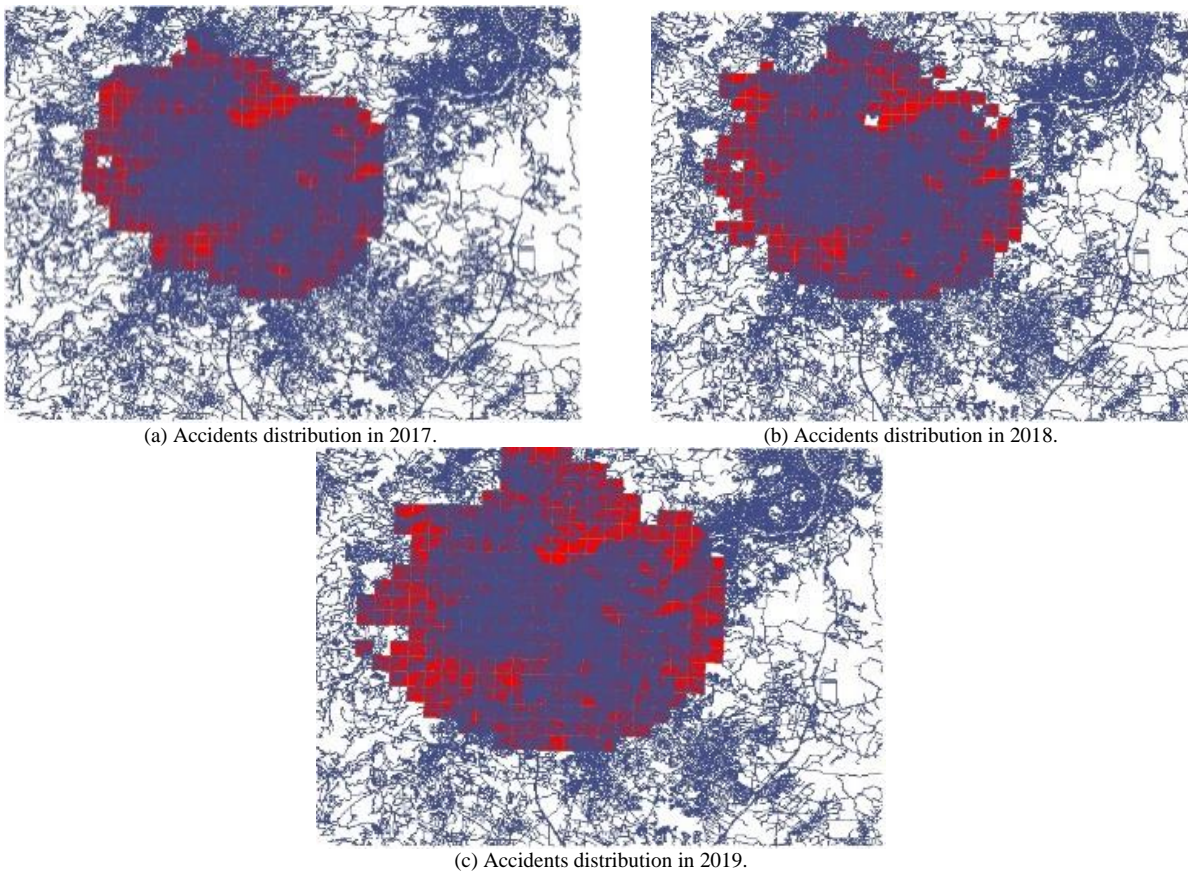


Fig. 8. Distribution of Hotspots of Accidents in 2017 (a), 2018 (b) and 2019 (c).

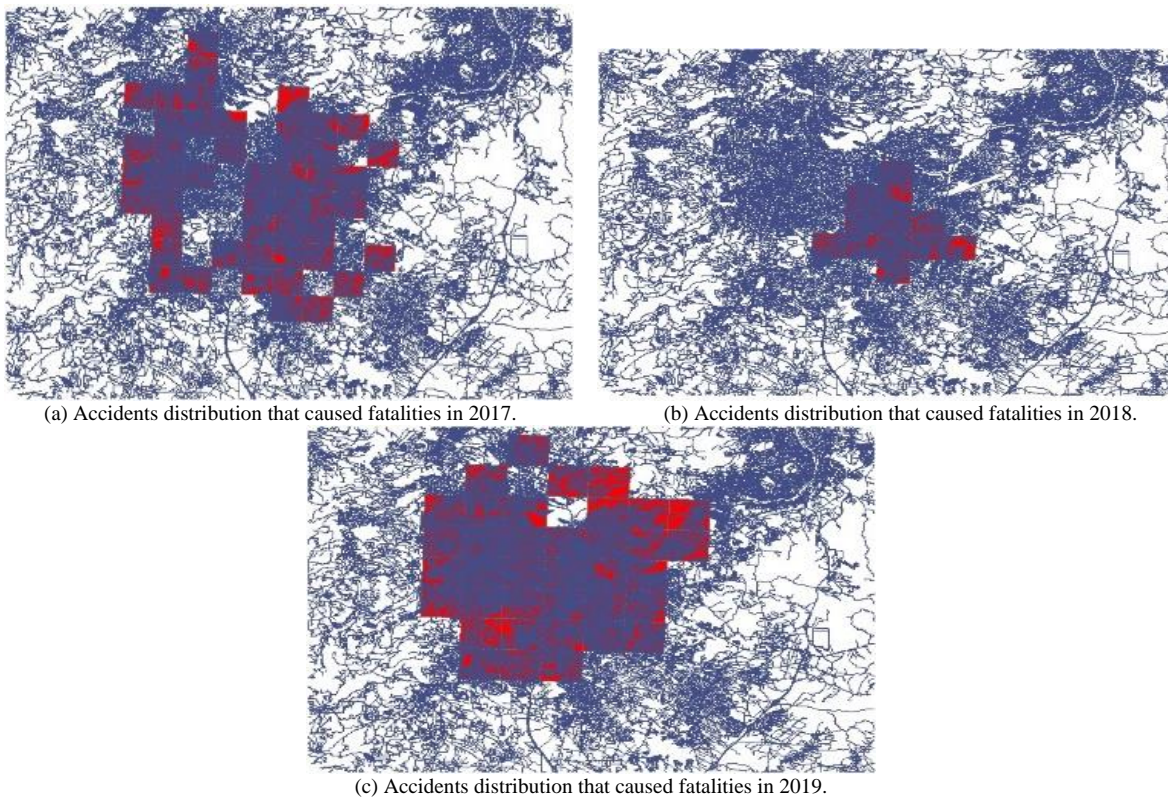


Fig. 9. Distribution of Hotspots of Accidents causing Fatalities in 2017 (a), 2018 (b) and 2019 (c).

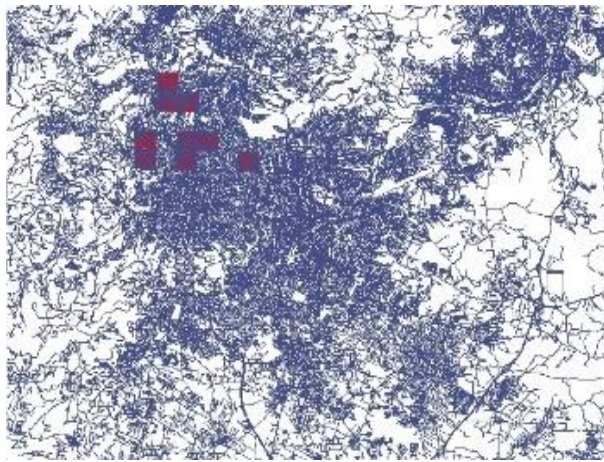
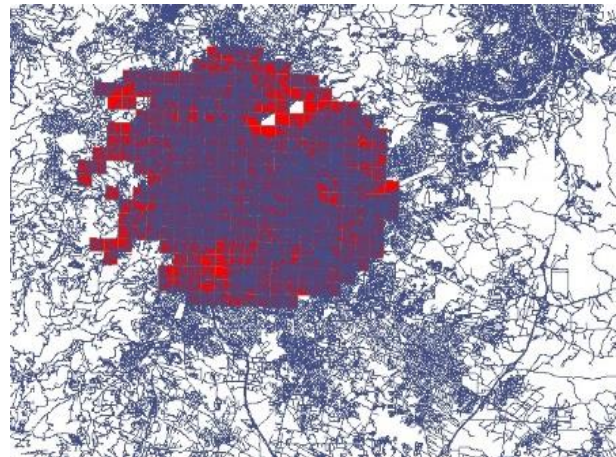


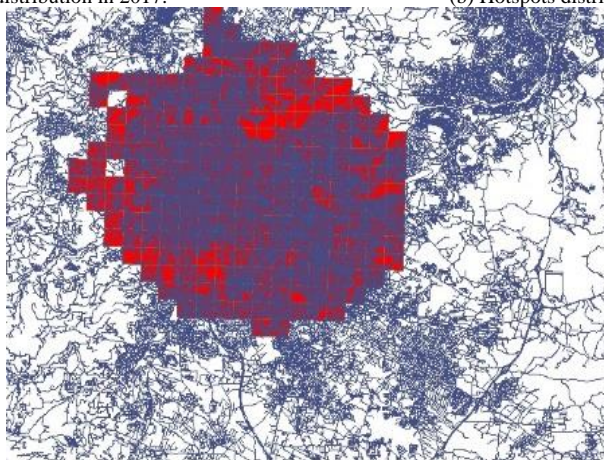
Fig. 10. The Distribution of Hotspots of Accidents that Caused Serious Injuries in 2018.



(a) Hotspots distribution in 2017.



(b) Hotspots distribution in 2018.

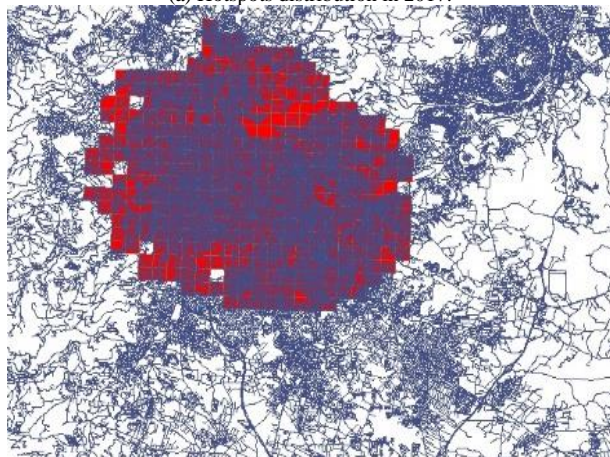


(c) Hotspots distribution in 2019.

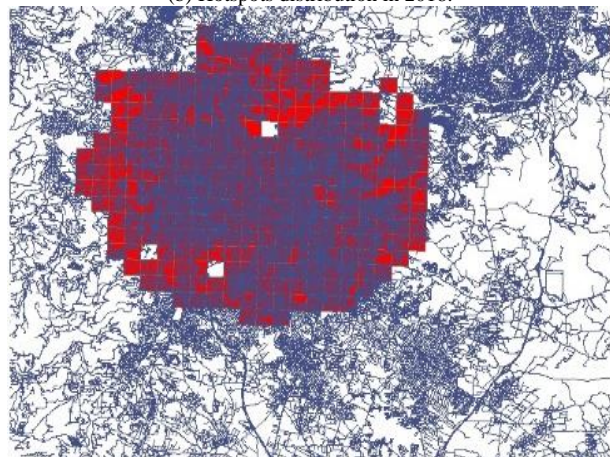
Fig. 11. Distribution of Hotspots of Accidents that Occurred on Roads Attached with Two Lanes Separated by a Central Island in 2017 (a), 2018 (b) and 2019 (c).



(a) Hotspots distribution in 2017.



(b) Hotspots distribution in 2018.



(c) Hotspots distribution in 2019.

Fig. 12. Distribution of Hotspots of Accidents that Occurred in the Road with Two Lanes not Separated by a Central Island in 2017, 2018 and 2019 from Left to Right respectively.

V. CRASH ANALYSIS OF TRANSPORT ROADS

One of the main objectives of this study is the analysis of safety along with various transport roads which are expected to gather more population (drivers and passengers). A new database schema was created to describe the collected data as shown in **Error! Reference source not found.** **Error! Reference source not found.** The parameters that are

considered are as follows: Street Name, number of accidents (#A), number of injuries (#I), number of fatalities (#F), and a status mode on a Public Transport Road (PTR).

TABLE III. HAZARDOUS STREETS IN AMMAN. WHERE #A, #I, #F AND PTR ARE THE NUMBERS OF ACCIDENTS, INJURIES, FATALITIES AND ON A TRANSPORT ROAD RESPECTIVELY

#	Street Name	#A	#I	#F	OTR
---	-------------	----	----	----	-----

The results show that 85.31% of accidents have occurred on roads of public transport. Also, the number of fatalities is increased by this type of road. The accidents are focused on the middle city of Amman. This indicates that the more cars, the more accidents. The roads with insufficient lights are not the main cause of accidents, the statistics have revealed that the lack of road lights has increased the drivers' attention while driving, leading to enforce them to reduce the speed. The recommendations of this research to the authorities of traffic control are: to set special lanes for the buses, enhance the highway road and provide more safety signals, assign special lanes for emergency vehicles such as ambulances and fire fighting, and assign CCTV cameras on the hotspots areas.

VI. CONCLUSION AND FUTURE WORKS

This study aimed to examine the accident data in the capital city Amman, Jordan and find the temporal and spatial distribution of these accidents and finally identify the hotspot in the study area.

SQL was used to perform the temporal analysis based on several categories. Results showed that accidents located on roads with two lanes are significantly more than accidents located on other types of roads. Moreover, this work aims to examine the accidents in the severity context and the accidents occurring each year, thus GIS-based tools were utilized to further examine these categories in a spatial framework.

Results highlight evidence of spatial clustering and recurrence of traffic accidents in the study area. Accident hotspots-based hazard maps for 2017 year accidents, 2018 year accidents, 2019 year accidents, two severity groups (Fatalities and serious injuries) and two road type groups (road of two lanes separated by central island and road of two lanes not separated by central island) were produced at the zonal level in the study area. The examined groups of accidents were found significant (>90%) for hotspots mapping.

Findings confirm that hotspots for all examined groups were mainly concentrated in commercial /residential and industrial/ zones which are located in and around the central regions of the study area. This is justified by that such regions have high density of population and public facilities, as a result high traffic volume and low speed limits.

More importantly, the findings contribute in elevating the systematic understanding of the spatial factor of traffic accidents in the study area and the influence of the geography and neighborhood of road in addition to other geometric features of roads on the traffic accidents. This could help in modelling and predicting high-risk zones and as a result taking into considerations precautions and developing suitable countermeasures for the identified hotspots zones.

It is worth mentioning that the accuracy of accident analysis depends on the accuracy of the original data. The accuracy of accident data can be improved by continuously update the records system used in the traffic department. This could be achieved by using Global Positioning System (GPS) receivers by the traffic police station to obtain accurate information about the locations impacted by the accident. The availability of such accurate information enables conducting efficient safety analysis. An accident diagram along with pattern analysis can be produced using special software.

Future work can be extended by taking into consideration more attributes of accidents such as reasons of accidents, identifying the relationship between traffic accidents and key the factors contributing to these accidents, and applying machine learning to predicate hotspots for other areas.

ACKNOWLEDGMENT

Wesam Alkhadour, Jamal Zraqou, Adnan Al-Helali, and Sajeda Al-Ghananeem are grateful to Isra University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

REFERENCES

- [1] Organization, W.H., "Global Status Report on Road Safety 2018," WHO: Geneva, Switzerland. 2019.
- [2] Wach, W., "Calculation reliability in vehicle accident reconstruction. Forensic science international," 2016. 263: p. 27-38.
- [3] Organization, W.H., "Global status report on road safety: Geneva.", 2018.
- [4] Statistics, D.o.. "Jordan Statistical Yearbook," Amman Jordan. 2018.
- [5] Al-Masaeid, H., "Jordan University of Science and Technology," Civil Engineering Dept, Irbid, Jordan, In a sabbatical leave at Aal Al-Bayt University, Al-Mafraq, Jordan). E-Mail: masaeid@aabu.edu.jo. 2018.
- [6] AECOM Consulting Transportation Group, Bellomo-McGee, and Ned Levine & Associates., "Considering Safety in the Transportation Planning Process," Federal, Highway Administration, U.S. Department of Transportation: Washington, DC. 2002.
- [7] Erdogan, S., et al., "Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar," Accident Analysis & Prevention, 2008. 40(1): p. 174-181.
- [8] Qu, X. and Q. Meng, "A note on hotspot identification for urban expressways. Safety Science," 2014. 66: p. 87-91.
- [9] Debrabant, B., et al., "Identifying traffic accident black spots with Poisson-Tweedie models," Accident Analysis & Prevention, 2018. 111: p. 147-154.
- [10] Ghadi, M., and Á. Török, "Comparison of different black spot identification methods," Transportation research procedia, 2017. 27: p. 1105-1112.
- [11] Rahman, M.T., A. Jamal, and H.M. Al-Ahmadi, "Examining Hotspots of Traffic Collisions and their Spatial Relationships with Land Use: A GIS-Based Geographically Weighted Regression Approach for Dammam," Saudi Arabia. ISPRS International Journal of Geo-Information, 2020. 9(9): p. 540.
- [12] Anderson, T.K., "Kernel density estimation and K-means clustering to profile road accident hotspots," Accident Analysis & Prevention, 2009. 41(3): p. 359-364.
- [13] Xie, Z. and J. Yan, "Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach," Journal of transport geography, 2013. 31: p. 64-71.
- [14] Moons, E., T. Brijs, and G. Wets, "Identifying hazardous road locations: hot spots versus hot zones," in Transactions on Computational Science VI. 2009, Springer. p. 288-300.
- [15] Shariff, S.S.R., et al., "Determining hotspots of road accidents using spatial analysis. Indonesia," J. Electr. Eng. Comput. Sci, 2018. 9(1): p. 146-151.
- [16] Loo, B.P., "Validating crash locations for quantitative spatial analysis: a GIS-based approach. Accident Analysis & Prevention," 2006. 38(5): p. 879-886.
- [17] Zahid, M., et al., "Predicting Risky and Aggressive Driving Behavior among Taxi Drivers: Do Spatio-Temporal Attributes Matter?," International Journal of Environmental Research and Public Health, 2020. 17(11): p. 3937.
- [18] Hirasawa, M. and M. Asano, "Development of traffic accident analysis system using GIS," 2001: Civil Engineering Research Institute.
- [19] Liang, L.Y., D.M. Ma'soem, and L.T. Hua, "Traffic accident application using geographic information system," Journal of the Eastern Asia Society for Transportation Studies, 2005. 6: p. 3574-3589.
- [20] Mali, S., "Traffic police operation based on sensors and data analytics," Transportation research procedia, 2020. 47: p. 187-194.
- [21] Pleerux, N., "Geographic Information System-based Analysis to Identify the Spatiotemporal Patterns of Road Accidents in Sri Racha," Chon Buri, Thailand. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 2020. 20(1): p. 59-70.
- [22] Feng, Y. and W. Zhu, "Formulating an Innovative Spatial-Autocorrelation-based Method for Identifying Road Accident Hot Zones," E&ES, 2020. 446(5): p. 052068.
- [23] Al-Omari, A., et al., "Prediction of traffic accidents hot spots using fuzzy logic and GIS," Applied Geomatics, 2020. 12(2): p. 149-161.
- [24] Yuan, T., X. Zeng, and T. Shi, "Identifying Urban Road Black Spots with a Novel Method Based on the Firefly Clustering Algorithm and a Geographic Information System," Sustainability, 2020. 12(5): p. 2091.
- [25] Pajaziti, A. and O. Tafilaj. "Traffic Accidents Analysis with the GPS/Arc/GIS Telecommunication System," Proceedings of SAI Intelligent Systems Conference. 2020. Springer.
- [26] Levine, N. "Building a spatial crash information system: Examples from the Houston-Galveston metropolitan safety planning," ITE 2006 Technical Conference and Exhibit Compendium of Technical Papers. 2006.
- [27] Levin, N. and N. Levine, "CrimeStat III-A spatial statistics program for the analysis of crime incident locations," US Department of Justice, Houston, 2004.
- [28] Amira K. Al-Aamri, Graeme Hornby, Li-Chun Zhang, Abdullah A. Al-Maniri, Sabu S. Padmadas, "Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman," Spatial Statistics, Volume 42, 2021, ISSN 2211-6753,
- [29] Kun Xie, Kaan Ozbay, Di Yang, Chuan Xu, Hong Yang, "Modeling bicycle crash costs using big data: A grid-cell-based Tobit model with random parameters," Journal of Transport Geography, Volume 91, 2021, ISSN 0966-6923.
- [30] Manap N, Borhan MN, Yazid MRM, Hambali MKA, Rohan A. "Identification of Hotspot Segments with a Risk of Heavy-Vehicle Accidents Based on Spatial Analysis at Controlled-Access Highway," Sustainability. 2021; 13(3):1487.
- [31] Obaidat, M.T., and T.M. Ramadan, "Traffic accidents at hazardous locations of urban roads," Jordan Journal of Civil Engineering, 2012. 159(700): p. 1-12.
- [32] Levine, N., "Considering Safety in the Transportation Planning Process," 2002.

An Agent-based Evaluation Model of Students' Emotional Engagement in Classroom

Moamin A. Mahmoud¹

Institute of Informatics and Computing in Energy (IICE)
Universiti Tenaga Nasional
Malaysia

Latha Subramainan²

College of Graduate Studies
Universiti Tenaga Nasional
Malaysia

Ihab L Hussein Alsammak³

Directorate General of Education of Karbala
Ministry of Education
Iraq

Mahmood H. Hussein⁴

Faculty of Computer Science and Information Technology
University of Malaya
Malaysia

Abstract—This study proposes an agent-based evaluation model of students' emotional engagement in a classroom. The proposed model consists of four main elements in a classroom which are, the selected strategy to control engagement, the engagement level of students, the emotional state of a lecturer, and the emotional states of students. The process starts with a lecturer selecting a strategy, which in turn influences the students' emotional state. By utilizing the three variables, students' misbehaviors, motivation, and participation, the engagement level of students is measured that eventually influences the lecturer's emotion either positively or negatively. If negatively, the lecturer proposes another strategy that would trigger the students' emotions and eventually improves the students' engagement level. We simulate our model to validate the applicability and functionality of the model. The simulation result shows a promising application to simulate a classroom environment with very flexible settings that leads to results in less time and cost. It also shows to be widely utilized by researchers in the field of social studies for further investigation of the problem of students' engagement by conducting experiments and report the results.

Keywords—Students' emotional engagement; agent-based evaluation; computational model

I. INTRODUCTION

Students' engagement in classrooms is a long-standing issue that needs to be objectively and cooperatively resolved or mitigated. It has always been regarded as a crucial factor that influences several educational outcomes [1]. Emotionally engaged students are highly motivated to attend classes, and actively participate in discussions and assessments [2]. On the other hand, emotionally disengaged students are usually habitual truants and/or occasionally misbehave in classrooms [3]. According to Yazzie-Mintz [4], studies have highlighted that boredom is a sign of lack of engagement during a lesson. For instance, bored students may apply less effort and stop paying attention to their lecturer, subsequently becoming even more bored over time and tend to do other unnecessary things. Such students ultimately end up getting poor academic results, involving in many disciplinary actions, and occasionally dropping out of colleges. As expected by the researchers, the

study found that teaching practices, teachers' and peers' support, and parental emotional support have a significant relationship with students' engagement. Research studies also reported that educators in many universities use a teacher-centered learning process strategy, which lacks personal autonomy, instead of a student-centered strategy [5] [6]. Asian students are notorious for their low-level class participation, it was revealed that less than 20% of students ask questions during class [7] [8]. These phenomena are due to their disinterest or boredom with ongoing lessons and are not concerned to participate in-class activities. Malaysia PISA survey measured students' engagement and found 80% of the participating schools fell into the poor performance bracket [9] [10]. Teaching strategies also have been seen as one of the support factors for students' engagement in schools [11].

Several research studies have been conducted to study the problem of poor engagement in classrooms using traditional methods such as questionnaire surveys, experimentation, and analysis. A case study was conducted to examine Malaysian school students' engagement status and to understand the factors that influence students' engagement in three psychological domains (behavior, emotional, and cognitive). Overall, the results show that the level of students' engagement is mostly influenced by emotion, followed by cognitive and behavior [30]. Studies by UCLA Higher Education Research Institute and British universities found that 40% of students are frequently bored in class due to poor engagement in classrooms [12] and 59% found lectures are boring in at least half of their classes [13]. Similarly, an annual survey by Indiana University's High School, discovered that about 30% of the students indicate that they are bored due to the lack of interaction with teachers and 75% report that the subjects being taught are not interesting [14]. Despite the widespread studies of student engagement in the higher learning institutions of Australia, the USA, and Canada, students' engagement research in Malaysian public universities is scant [15]. However, existing work on student engagement research suffers from the following deficiencies:

1) Many of the experiments conducted by social science researchers used traditional methods that consume more time and effort. Furthermore, since these experiments are applied to humans, there are very limited settings to be tested and the costs would be excessive if multiple settings are applied [16]. Studies have indicated that it is challenging to manage and control humans to conduct multiple experiments or repeat experiments with the same settings [17]. Thus, traditional methods have a poor success rate in mitigating the problem.

2) Social studies have identified three dimensions of student engagement which are Behavioural Cognitive, and Emotional engagement [18]. While ample research has been conducted on Behavioural and Cognitive engagement, Emotional engagement has received little attention [19]. This is attested by Pekrun et al. [20] who surmises that research on emotion has mostly been neglected as a factor to improve students' engagement although researchers from social studies emphasize the importance of emotion on students' engagement.

3) On the other hand, when it comes to emotional disengagement, a survey of student engagement reported student felt so bored in class, 81% of the students responded that the material wasn't interesting and 42% of the students felt that lack of relevance caused boredom [21]. The perceived uninteresting teaching strategy is the biggest contributing factor that needs to be improved upon. However, the development of interesting strategies takes a great deal of time and effort. Besides that, there are tons of strategies that have been proposed by many researchers. Unfortunately, the researchers do not provide the applicability of the strategies in any domain. Moreover, there is no proper centralized library to hold a collection of strategies as the strategies are scattered on the Internet. It is essential to design strategies that help students connect with learning and improve engagement [31] [32] [33].

To overcome these limitations, we propose a study to formulate an agent-based evaluation model of students' emotional engagement in a classroom. In this research, software agents will be empowered to simulate the different states of students' emotions to animate a scenario that reflects or simulates a real classroom environment [27] [28] [29]. To do so, this work primarily focuses on student engagement and its possible classroom teaching strategies to improve students' negative emotional states and engagement performance. We discuss several factors that have been influencing the student engagement process in the literature. However, we emphasize emotions as a factor to shape student engagement as research on emotions in education has received a great deal of attention among educational psychologists [22] [34] [35]. To complete the scope of the engagement, we identify the common emotional experiences of students and lecturer that takes places in a classroom that reflect the engagement level. We then proposed a method to measure the engagement level in a classroom by utilizing three engagement factors [36] [37]. On the other hand, to propose possible strategies to cater to any of the negative emotional states of students, we identify two sets

of attributes related to this study, environmental and emotional attributes through the literature. Environmental attributes include a number of students (30 or 60 or 90 or 120), class session (e.g. morning, afternoon/evening), class duration (1 hour or 2 hours or more), type of subject (e.g. theoretical or conceptual), year of study (junior or senior), lecturer style (controlling or autonomy-supportive) and emotional attributes are those such as negative emotional states of students (e.g. anger, anxiety or boredom). Finally, we create a virtual environment simulating classroom dynamics by animating a lecturer agent. The lecturer can insert any strategy using our strategy specification setting and test it in an environment setting to observe the success rate of each strategy to improve student engagement based on their classroom environmental factors. Figure 1 shows the scope of this thesis.

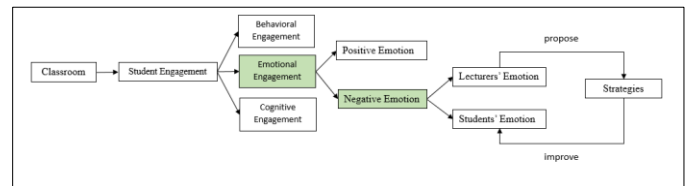


Fig. 1. Research Scope.

The outcomes of this study include an agent-based evaluation model, a method to evaluate emotional engagement level, identify emotional states of students, and propose the most suitable teaching strategies with strategy strength and a classroom dynamics simulator. Researchers and educationists could utilize the simulator to investigate the problem of students' engagement. This could benefit schools and universities by giving teachers and lecturers exposure to students' engagement for improved academic performance.

II. AGENT-BASED EVALUATION MODEL

Typically, a lecturer monitor student's engagement level in a classroom via indirect indicators, for instance, amount of participation in classroom discussion, attendance, commitment on a task given, time spent on assessments, the intensity of concentration during the ongoing lesson, and amount of motivation or interest shown on particular course material Lamborn, (1992). The preliminary model of this study is based on four main elements in a classroom; the selected strategy to control engagement, the engagement level of students, the emotional state of students, and lecturer emotion. An applied strategy during a classroom session influences the students' engagement-based emotions. By utilizing three variables, students' misbehavior, poor motivation, and poor participation, the engagement level of students can be measured. Student's emotional state can be identified through analysis of engagement level. A negative student's emotional state and the result of engagement measurement trigger the lecturer's emotion either positively or negatively. If affected negatively, the lecturer proposes another strategy that would trigger students' emotions positively and eventually improve engagement. The lecturer again measures the engagement level and decides whether to take another strategy or maintain the existing one. Figure 2 illustrates the agent-based evaluation model.

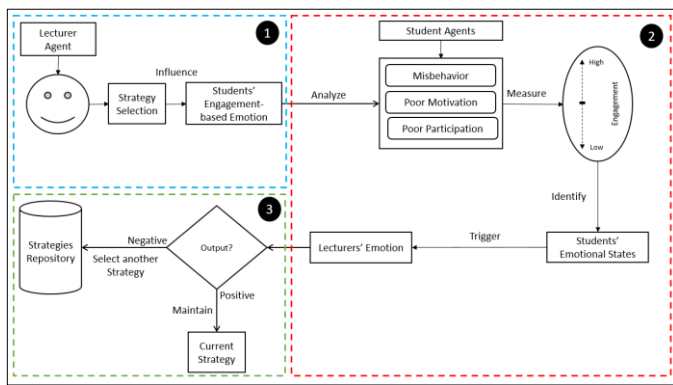


Fig. 2. Agent-based Evaluation Model.

As shown in Figure 2, the classroom simulator model is divided into three main components. The first component shows the Lecturer Agent (AL) Components that involve a strategy element. The AL applies the strategy during the classroom session. The effectiveness of the strategy applied influences students' engagement-based emotions. The second component shows the Engagement Measurement by analyzing the factors of misbehavior, poor motivation, and participation. Students' emotional state can be predicted through the result of engagement. Thus, negative emotional states of students and the level of engagement affect the AL emotions which trigger either positively or negatively. The third component shows the Strategy Revision where the negative emotion of the lecturer triggers the AL to select a new strategy that would improve their engagement in the classroom. The AL maintains the current strategy if positive emotion occurred. The process continues until the student displays positive factors of behavior, a high level of motivation, and participation. The logical model has been presented in each component in the next section.

A. First Component - Lecturer Agent and Strategy Selection

1) *Modelling lecturer agents' architecture:* In the Belief-Desire-Intention (BDI) architecture [26], agents are built with mentalistic notions to recognize or deliberate their goals. An agent's belief represents its knowledge about the environment. It forms its belief from the state of the environment. If there are changes in the state of the environment, it updates its belief, which changes its behavior that leads to the achievement of its goal. The Emotional component represents the emotions being experienced by the lecturer towards students' engagement in a classroom medium, high, or very high influenced by strategy application. The Belief component is associated with the Desire component aims to achieve an acceptable engagement level by revising or maintain a strategy determined by the Intentional component. These emotions trigger the Lecturer's Belief about the current engagement level that is either very low or low. The agent belief will be triggered when the LA receives or observes a negative signal(s) of students' engagement. When the Belief is triggered (e.g. feels worried or upset), the LA Desire will be revising the current strategy based on the received emotion to enhance the engagement. Once the LA figured out a better

strategy, the intention will be to implement the revised strategy.

2) *Strategy selection:* It begins with strategy selection by the lecturer. As explained earlier, a few factors have been influencing classroom strategy. In our study, we only emphasize class environment factors such as the number of students (30 or 60 or 90), class session (eg. morning, afternoon/evening), class duration (1hour or 2 hours or more), type of subject (e.g. theoretical or conceptual), and year of the student (junior or senior). Once the lecturer entered the environmental setting data, the next step is to evaluate student's behavior (engagement data) based on previous classroom sessions and finally select a strategy from the drop-down list. The effectiveness of any strategy applied in turn influences students' engagement-based emotion.

3) *Logical model of strategy selection:* Once the lecturer selects a strategy and runs the environment setting, the AL capture the environment setting data and test the effectiveness of the selected strategy. AL analyses the data inserted and produce the strength of the selected strategy in percentages. Besides that, AL can cross-find the nearest matched strategies with environment and engagement data entered by the lecturer. Figure 3 shows the strategy, selection model. We define the terms Environment Factors and Engagement Factors as follows:

Definition 1: Environmental factors, EnvFact consist of Number of Students, Class Session, Class Duration, Type of Subject, and Year of Study.

Definition 2: Engagement factors, EngmtFact constitute of Misbehavior, MisB, Poor Motivation, Mtv, and Poor Participation, Ptc.

Definition 3: A Strategy Selection, StrgySelect constitutes of Environmental Factors and Engagement Factors.

StrgySelect: { EnvFact, EngmtFact }

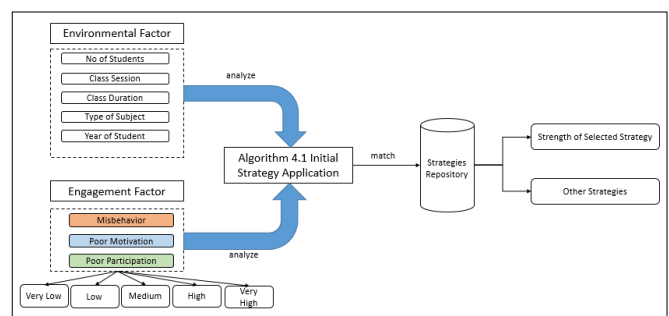


Fig. 3. Strategy Selection Model.

There are various classroom environmental factors have to be considered before a lecturer or social studies researchers can identify the best teaching strategy to be applied to keep student remain engaged during the lesson. For example, they need to consider the number of students, a class session with the duration of the lesson, the type of subject to be taught, and the seniority of students. The way students behaving, level of motivation, and amount of participation during a lesson reflects

how much they are engaged with the lesson. By considering these factors, a lecturer would have a clear idea of a suitable teaching strategy to be applied. Using this tool, the lecturer can find out the strength of the selected teaching strategy according to the environmental and engagement factors. However, if he/she made a poor selection teaching strategy, the agent proposes the best three strategies.

B. Second Component - Engagement Measurement

As explained in Section 1, the most prominent emotional states that occurred in a classroom include enjoyment, pride, anger, anxiety, and boredom. For example, before a classroom lesson start, less prepared or low self-efficacious students are more likely to experience negative emotions. In our context, anger, anxiety, and boredom represent the states of negative emotions. In contrast, well-prepared or highly self-efficacious students are evaluated to be in a state of positive emotional state. Enjoyment and pride emotions represent the elicited positive emotional state for our context at the beginning of the lesson. By comparison, students are expected to experience different elicited emotional state after the lecturer change their teaching strategy. At this stage, different learning strategies give rise to elicited emotions. For example, a desirable strategy (e.g. engaged in peer review) elicits a state of positive emotions. In contrast, an undesirable strategy (e.g. reading slide contents) elicits a state of negative emotions. However, literature reported that a lecturer needs to change their teaching strategies only when they observed the students are experiencing negative emotional states. To evaluate the student's emotional state, we need to identify the values of the engagement factors (e.g. Misbehaviour, Poor Motivation, and Poor Participation) that have been influencing the engagement level in a classroom and subsequently the lecturer's emotions. These are the potential factors to infer students' engagement levels. Generally, negative emotional states will impact engagement level, however, we need to identify the intensity of negative emotions influencing the students in a classroom. This is because several emotions can occur at the same time among different individuals. Changes in the emotional states of students consequently reflect the engagement level in a classroom. Therefore, several student behavior indications reveal the emotional states of students. It is a well-known fact from many studies that indicates a positive emotional state leads to high-level engagement in the classroom including no misbehaving attitude from a student, a high level of motivation, and good participation in class discussion. Whereas, negative emotional states will lead to low engagement whereby students pay no attention and started to do their work with no motivation and participation.

1) *Engagement level*: In the past, many researcher proposed techniques to measure the engagement, and its' have been practiced in a different context to identify if students are actively engaged in the learning process. These measures emphasize the traditionally "quantifiable" aspects of attendance rates, truancy, time-on-task, and consequently suspension/discipline rates [23]. Besides quantifiable methods, self-report, teacher ratings, interviews, observations, cross-cultural data, and assessment grades also have been in practice. Despite the traditional method of measuring

engagement, researchers stressed that more systematic and thoughtful attention to the measurement of engagement is the most imperative direction for future research. [24].

Therefore, in our study, we anticipate modeling how student engagement can be measured based on the variable of engagement factors; misbehavior, motivation, and participation of students. We use six-level scales; not present, very low, low, medium, high, and very high to rate student behaviors from various engagement factors parameters. To determine the current engagement level, the lecturer is required to rate student behavior in percentages based on their interaction with their student in the last class session. Therefore, they are required to use six-level scales as shown in Table 1 to determine the student behaviors. In this section, we use the ordinal scale by Stevens (1946) to determine the engagement level as follows in Table 1.

The process starts with measuring the engagement level of students by the lecturer via indicators of three variables: students' misbehaviors, poor motivation, and poor participation. We have a total of 11 indicators for the 3 engagement factors (eg. Misbehavior (4 indicators), Poor Motivation (4 indicators), and Poor Participation (3 indicators)). Once we analyzed the value of engagement factors, we can identify the current engagement level in a classroom. To measure the engagement level of students from various factors, we use the ordinal scale as per Table 1. Figure 4 shows the engagement level analysis model. We define the terms Engagement Level as follows:

Definition 4: Engagement level, EngLvl can be measured based on average values of engagement factors; MisB, Mtv, and Ptc level in the classroom.

2) *Modeling student agents*: The students' agents in this classroom simulator tools are very basic. This agent reflects the behavior as indicated by the academician. For example, let's denote the number of a student as agent α_n , misbehavior indicators as X_n , poor motivation indicators as Y_n , poor participation as Z_n , and rating values as values (e.g. 0 if not present and 1 if present, if the rating falls into any level of scales), Table 2 shows the indicators of each behavior.

We use the randomize function to randomly assign values to student agents. For instance, student behavior ($\alpha_1, X_3, 1$). Table 3 shows an example of evaluating student behavior for a class of ten students.

TABLE I. SIX LEVEL SCALES

Not Present	Very Low	Low	Medium	High	Very High
0	1 - 20	21 - 40	41 - 60	61 - 80	81 - 100

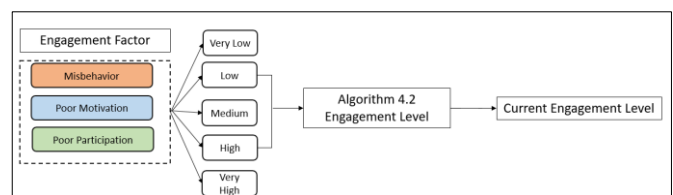


Fig. 4. Engagement Level Analysis Model.

TABLE II. INDICATORS OF STUDENTS' BEHAVIOR

Four indicators describe <i>Misbehaving</i>	Verbally or physically aggressive	X1
	Spoken in a raised voice	X2
	Acting in an abusive manner	X3
	Yelling and screaming	X4
Four indicators describe poor <i>Motivation</i>	Non-attentiveness	Y1
	Talking out of turns	Y2
	Playing video games	Y3
	Get away from class	Y4
Three indicators describe the poor <i>participation</i>	Avoiding eye contact	Z1
	Inability to initiate conversations	Z2
	Avoidance or refusal to participate	Z3

TABLE III. EXAMPLE OF EVALUATING STUDENT BEHAVIOR

	X ₁	X ₂	X ₃	X ₄	Y ₅	Y ₆	Y ₇	Y ₈	Z ₉	Z ₁₀	Z ₁₁	
	Misbehavior				Poor Motivation				Poor Participation			
α ₁	0	0	0	0	0	0	0	0	0	0	0	-
α ₂	1	0	0	0	1	0	0	0	0	1	0	√
α ₃	0	1	0	0	0	0	0	0	0	0	0	√
α ₄	1	0	0	0	0	0	0	1	0	0	1	√
α ₅	0	0	1	0	0	1	0	0	1	0	0	√
α ₆	0	0	0	1	1	0	0	0	0	0	0	√
α ₇	0	1	0	0	0	0	0	0	0	1	0	√
α ₈	1	0	0	0	1	1	0	0	0	0	0	√
α ₉	0	0	0	0	0	0	0	0	0	0	0	-
α ₁₀	0	0	0	0	0	0	0	0	0	0	0	-
	7				5				4			
Total number of students who are not engaged to lesson												7

By analyzing Table 3, 7 out of 10 students are not engaged in the lesson, only 3 of them are engaged in the lesson. Therefore, current Engagement Level is $(3 / 10) * 100 = 30\%$. According to our ordinal scales in Table 1, the current engagement level is LOW. However, the total number of students in the classroom has an impact on the result of the engagement level. This is because students experience more than one type of emotional state at one time.

3) *Identify the emotional states of student:* The AL can identify the emotional states of students based on the values of engagement factors (refer to Table 3). In this section, we use the ordinal scale by Hogan and Warrenfeltz [25] to determine the scale of the outcomes for each factor in Table 2. Once we analyzed the value of engagement factors, we can identify the intensity of the emotional states of students. Figure 5 shows the identification of the intensity of the emotional state's model. However, to conclude the values of the emotional factors, we use the ordinal scale by Hogan and Warrenfeltz [25] to determine the scale of the outcomes for each factor in Table 4.

TABLE IV. ORDINAL SCALE [25]

LOW	MEDIUM	HIGH
10% - 39%	40% - 69%	70% - 100%

We define the terms Emotional Factors as follows:

Definition 5: Emotional factors, EmoFact constitute of Anger, Boredom, and Anxiety derived from EngmtFact.

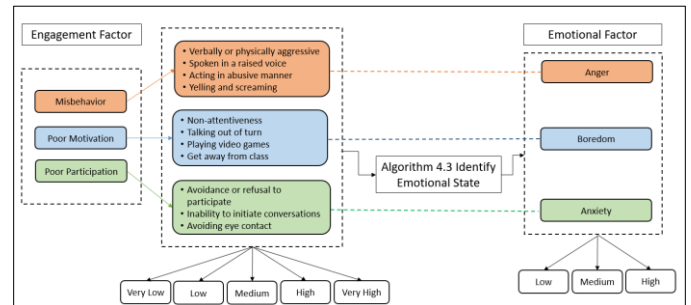


Fig. 5. Identification of Intensity of Students' Emotional States Model.

Therefore, according to Table 4

a) Total number of Students' Misbehavior is 7, therefore $(7 / 10) * 100 = 70\%$ - High.

b) Total number of Students' Poor Motivation is 5, therefore $(5 / 10) * 100 = 50\%$ Medium.

c) Total number of Students' Poor Participation is 4, therefore $(4 / 10) * 100 = 40\%$ Medium.

As a result, the intensity of the emotional states of students as follows:

EmoFact (MisB) = (Anger, High)

EmoFact(Mtv) = (Boredom, Medium)

EmoFact (Ptc) = (Anxiety, Medium)

The lecturer measures the engagement level of students by rating the students' behaviors based on past experiences. From engagement factors result, The agent can derive the intensity (eg. low, medium, high) of emotional states of students; anger, boredom, and anxiety.

4) *Identify lecturer emotion:* Lecturer's emotions are as important as students' emotions. The result of the engagement measurement positively or negatively affects a lecturer's emotion. If negatively, the lecturer deploys another strategy that triggers positive students' emotions of which would eventually improve engagement. In contrast, the positively affected teacher is expected to engage a strategy that would maintain a positive aura to his students. According to literature, student emotional state also impacts on lecturer emotion. This engagement assessment cycle should continue to maintain positivity throughout the students learning sessions. Figure 6 shows the lecturer's emotion model. We define the terms Lecturer emotion as follows:

Definition 6: Lecturer emotion, EmLec is being affected by the emotional state of student Emotional Factor, EmoFact, and Engagement Level, EngLevel.

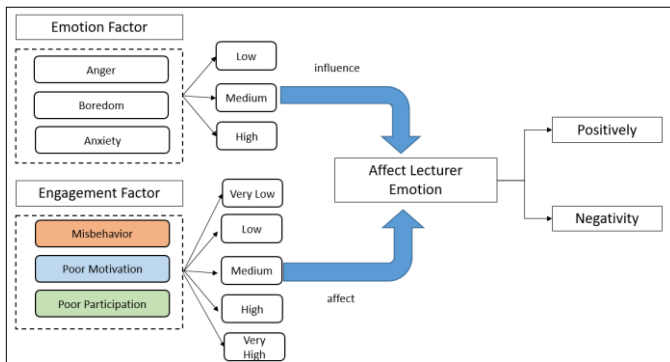


Fig. 6. Lecturer Emotion Model.

A lecturer is always concern about the engagement level of students during a lesson that directly impacts their emotions. It impacts the lecturer’s emotions either positively or negatively. For example, low engagement levels affect the emotion of the lecturer negatively and vice versa. If she/he is affected negatively, it would trigger them to change their teaching strategy to attain back their students’ attention. However, in this thesis, we focus formulation of identifying students’ emotional states only. Lecturer emotions are stated in the agent-based evaluation model to complete the whole cycle of enhancing the student engagement process. As per in the literature, students’ emotional state equally impacts the lecturer's emotions either positively or negatively that eventually triggers the lecturer to deploy another teaching strategy to keep students remain engaged with the lesson. Therefore, we do not have any formulation to identify lecturer emotions in this classroom simulator tool as we focus on identifying the students’ emotional states. Moreover, the lecturer agent can measure the engagement and identify the emotional states of students however identify or improving lecturer emotions is not our scope of the study.

C. Third Component - Strategy Revision

Once the lecture emotion is being affected negatively, the lecturer can implement other strategies as been proposed by the AL, then test the effectiveness of the new strategy again. Besides that, a lecturer, who might have ideas on new strategies based on their teaching experience, can insert a new strategy using the Strategy Specification Settings Interface and use the Likert scale to appropriately assign the impact of the strategy. The interface consists of a new strategy text field, a number of students (30 or 60 or 90), class session (eg. morning, afternoon/evening), class duration (1hour or 2 hours or more), type of subject (eg. theoretical or conceptual), and years of a student (junior or senior) in the drop-down list. For example, a lecturer from the Department of software engineering would like to register a new strategy that he had tried in his classroom and found he can improve the engagement level from low to high. However, he needs to indicate the class environment factors that are influence his success rate of strategy. For instance, his strategy for software testing theory subject only works for 60 third-year students, with a 2-hour duration at the morning session. He also needs to indicate the intensity of emotional states that he plans to improve (e.g high, medium, and low). Figure 7 shows the new

strategy storing model. We define the NewStrategy Application terms as follows:

Definition 7: NewStrategy application to be stored in Strategy Repository, StrgyRepo constitutes of Environmental Factors, EnvFact and Emotional Factors, EmoFact.

Once the lecturer filled up all the information requested, the strategy will be are stored in a repository as text files. Each new strategy application will be saved as a file. Later on, if the lecturer intends to edit the strategy, he/she can do so.

1) Proposing best three strategies: In the environment specification setting, it begins with strategy selection where the lecturer needs to key in information of her/his class environment and estimates the percentage of student behaviors based on their last class session experience which eventually predicts the students’ emotional states and engagement level. Then, they are required to select a strategy from the list. As output, the AL will display the engagement level and emotional states of students. The strength of the selected strategy in percentage will be displayed too. Besides, the agent can propose other strategies based on the environmental and engagement data inserted by the lecturer. Meanwhile, in a strategy specification setting, a lecturer can store new strategies that include environmental and emotional factors. Therefore, using two settings information, the AL can propose the best three strategies that suit the environment setting data entered by the lecturer. We define the terms environment setting and strategy specification setting as follows:

Definition 8: Environment Setting, EnvSet comprises of EnvFact, EngmFact, EmoFact, EngLevel, StrgySelect, and EmLec. Strategy Specification Setting, StrgySpecSet comprises of StrgyRepo.

Definition 9: Strategies Strength, Strgy Strength comprises EnvSet and StrgySpecSet

$$\text{Strgyn Strength: } \{ \text{EnvSet} \cap \text{StrgySpecSet} \}$$

Therefore to propose best 3 strategies:

Strategy Selection: MAX (Strgy1 Strength, Strgy2 Strength, Strgy3 Strength... n++).

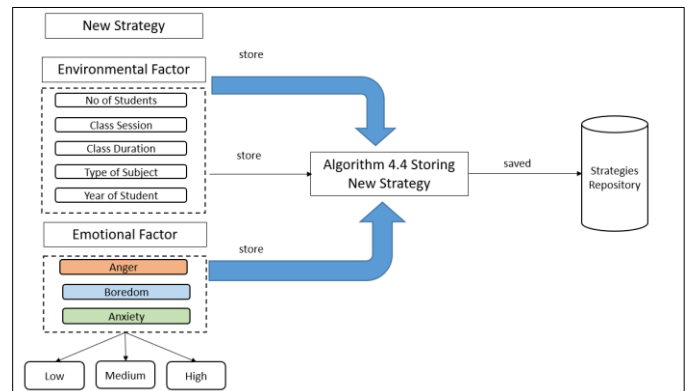


Fig. 7. New Strategy Storing Model.

However, to find the best 3 strategies' effectiveness, we need to analyze the Strgy Strength data. We have a total of 8 factors as follows:

a) Environmental Factors (5) - Number of Students, Class Session, Class Duration, Type of Subject, and Year of Study.

b) Emotional Factors (3) – Anger, Boredom, and Anxiety.

Hence, we will find the average of each scenario then divide it by 8 then convert it into percentages.

In this case scenario, the lecturer chose the Online Quiz strategy for his 2 hours, morning session class with 30 senior students, for his Software Quality theoretical subject. However, according to our simulator, the Online Quiz strategy does not apply to his environmental setting as the strategy strength is 12.5%. In other words, deploying this online quiz strategy for his class situation will lead to low engagement as students already experiencing medium-level anger and boredom, and high anxiety. This emotional feedback revealed as some of the students might not ready for an online quiz as no early preparation was done and some might be bored to answer questions so they might not pay attention to the questions. Therefore, the AL can propose other better strategies suitable for his environmental setting. In this case, Peer Review is the most suitable strategy to be applied for morning session classes for senior students as it has 87.5% strength followed by applying Real Life Scenarios strategy which is 62.5%, and Discovery and Discussion strength is 50%. In conclusion, by deploying a Peer Review strategy, the academician can cater to medium-level anger, and a high level of boredom and anxiety students and thus improve their emotional states and engagement level.

III. DEVELOPING THE AGENT-BASED EVALUATION MODEL USING PROLOG

We have chosen Prolog to simulate an agent-based evaluation model to create a real-world implementation for the classroom domain. We use Win-Prolog and its extended module Chimera, which can handle multi-agent systems. We use Prolog for two reasons: firstly, Prolog is well suited for expressing complex ideas because it focuses on the computation's logic rather than its mechanics where the drudgery of memory allocation, stack pointers, and computational engine. Secondly, since Prolog incorporates a logical inferencing mechanism, this powerful property can be exploited to develop inference engines specific to a particular domain.

A. Process Flow of Classroom Simulator Tool

The simulator is divided into two main settings which are Environment Specification Setting and Strategy Specification Setting. The main use for these two main settings is the academician and social studies researchers. In Environment Setting, the academician or the researcher can test the strategy selected whereas in Strategy Specification Setting, the lecturer can store their ideas of strategy, and then they can test it on Environment Setting to observe how much it can improve the student engagement performance in a classroom. Humans as Lecturer agents (AL) communicate with their agents via an

interface and the agents monitor and update their environment to communicate between agents, perform tasks that enable the progression of the workflow. For an initial strategy application, it constitutes environmental factors and engagement factors. From the engagement factors, our AL can derive the engagement level (e.g. very low, low, medium, high, very high) and emotional factors which include the negative emotional states of the student (e.g. anger, anxiety, or boredom).

A teaching strategy is applied based on a set of environment and engagement data, AL will analyze the data inserted by the lecturer. AL then measures the engagement level in the classroom by analyzing the engagement factors and emotional state of students can be predicted. In this case point, the ineffectiveness of a strategy can cause boredom among students thus will lead to low engagement. Subsequently, low engagement and negative emotional state of students affect lecturers' emotions. If affected positively, the lecturer can maintain the strategy, otherwise, implement another most suitable strategy. AL can propose other better strategies. In this simulator, a lecturer, who might have ideas on new strategies based on their teaching experience can store a new strategy application into Strategies Repository constitutes of Environmental Factors and Emotional Factors using a proposed New Strategy Specification Settings Interface. The core modules making up the agent-based evaluation model are the selected strategy to control engagement, the engagement level of students, and the emotional states of students in a classroom. Figure 8 shows the process flow of the classroom simulator tool.

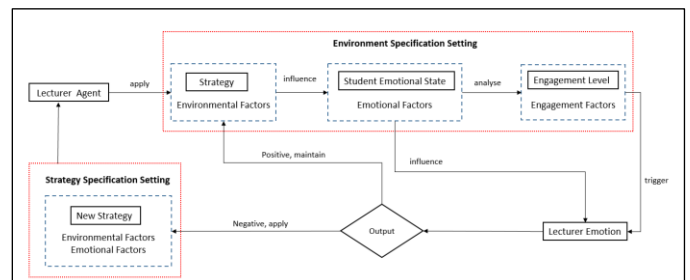


Fig. 8. Process Flow of Classroom Simulator.

B. Scenario-Based Testing

This section will be divided into two. In the first part, Environmental testing is conducted and in the second part, Strategy Specification testing is conducted. We conduct four scenario-based testing and two new strategies testing to validate our classroom simulator tool to find out the strategy strength and ability of agents to propose the three best strategies according to the environmental setting factors.

1) Environmental setting testing

a) Scenario A: Lecturer A would like to deploy a strategy for the “Software Testing” theory-based subject which is scheduled on Monday morning 2 hours class session which consists of 90 senior students. the rates of student behavior are assumed as follows in Table 5.

From the analysis of results in scenario A in Table 5, it shows that engagement level was very low in past class

sessions. Students experiencing high boredom followed by medium level anger and anxiety. In short, students' poor motivation is a high and medium state of misbehavior and participation among students. Selected strategy result and best 3 strategies proposed by agents as follows in Figure 9.

Analyses of results show that the initial strategy selection which is Discovery and Discussion accordance with the classroom environment is only 50% applicable. Agent proposed Peer Review would be a better strategy to deploy to improve poor engagement and emotional states of students in scenario A as the strength is 75% which is higher than Discovery and Discussion strategy strength followed by Real Life Scenario which is 62.5%.

b) Scenario B: Lecturer B would like to deploy a strategy for the "Discrete Structure" practical-based subject which is scheduled on Friday afternoon 2 hours class session which consists of 60 junior students. Based on his last class session, he rates student behavior as follows in Table 6.

From the analysis of the result in Table 7 accordance with scenario B in Table 7, shows that engagement level was very low in past class sessions. Students experiencing a high level of boredom and anxiety meanwhile anger emotional states do not present. In short, students' poor motivation and participation are high among students. Selected strategy result and best 3 strategies proposed by agents as follows in Figure 10.

TABLE V. RATING STUDENT BEHAVIOR FOR SCENARIO A

Engagement Factor	Indicators	Rating
Misbehavior	Verbally or physically aggressive	Low
	Spoken in a raised voice	High
	Acting in an abusive manner	Not Present
	Yelling and screaming	Not Present
Poor motivation	Non-attentiveness	Very High
	Talking out of turns	High
	Playing video games	Not Present
	Get away from class	High
Poor Participation	Avoiding eye contact	Medium
	Inability to initiate conversations	Medium
	Avoidance or refusal to participate	High

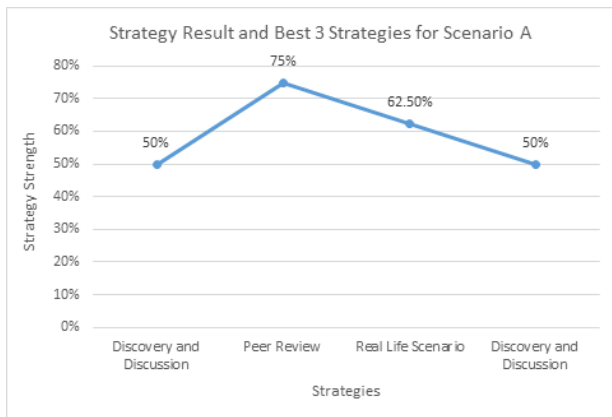


Fig. 9. Selected Strategy Result and Best 3 Strategies for Scenario A.

TABLE VI. RATING STUDENT BEHAVIOR FOR SCENARIO B

Engagement Factor	Indicators	Rating
Misbehavior	Verbally or physically aggressive	Not Present
	Spoken in a raised voice	Not Present
	Acting in an abusive manner	Not Present
	Yelling and screaming	Not Present
Poor motivation	Non-attentiveness	High
	Talking out of turns	Medium
	Playing video games	Low
	Get away from class	Low
Poor Participation	Avoiding eye contact	High
	Inability to initiate conversations	Low
	Avoidance or refusal to participate	High



Fig. 10. Selected Strategy Result and Best 3 Strategies for Scenario B.

Analyses of results show that the initial strategy selection which is Peer Review in accordance to the classroom environment is only 37.5% which is not suitable to deploy. Agent proposed Discovery and Discussion would be a better strategy to deploy to improve poor engagement and emotional states of students in scenario B as the strength is 75% which is higher than Peer Review strategy strength followed by Real Life Scenario which is 62.5% and Online Quiz which is 50%.

TABLE VII. RATING STUDENT BEHAVIOR FOR SCENARIO C

Engagement Factor	Indicators	Rating
Misbehavior	Verbally or physically aggressive	Very Low
	Spoken in a raised voice	Not present
	Acting in an abusive manner	Not present
	Yelling and screaming	Not Present
Poor motivation	Non-attentiveness	Medium
	Talking out of turns	High
	Playing video games	Medium
	Get away from class	Medium
Poor Participation	Avoiding eye contact	Medium
	Inability to initiate conversations	Low
	Avoidance or refusal to participate	Medium

c) *Scenario C*: Lecturer C would like to deploy a strategy for the “Operating System Concepts” theory-based subject which is scheduled on Thursday morning 1-hour class session which consists of 30 senior students. Based on his last class session, he rates student behavior as follows in Table 7.

The analysis of results in Table 6, it shows that the engagement level was low in past class sessions. Students experiencing a high level of boredom, followed by a medium level of anxiety and a low level of anger. In short, students’ poor motivation is a high and medium state of poor participation and a low level of misbehaving among students. Selected strategy result and best 3 strategies proposed by agents as follows in Figure 11.

Analyses of results show that the initial strategy selection which is Online Quiz accordance with the classroom environment is only 50% applicable. Agent proposed both Peer Review and Online Quiz strategies are the best strategies to deploy to improve the poor engagement and emotional states of students in scenario C as the strength is 50% followed by Online Assessment which is only 37.5%. For this scenario C, better strategies need to be stored in a tool that has higher strength than both Peer Review and Online Quiz strength.

d) *Scenario D*: Lecturer D would like to deploy a strategy for the “Java Programming” practical-based subject which is scheduled on Tuesday afternoon 2-hour class session which consists of 120 senior students. Based on his last class session, he rates student behavior as follows in Table 8.

The analysis of the result shows that the engagement level was medium in the last class session. Students experiencing a medium level of anger, boredom, and anxiety. In short, a medium state of students’ misbehavior, poor motivation, and participation among students. Selected strategy result and best 3 strategies proposed by agents as follows in Figure 12.

Analyses of results show that the initial strategy selection which is Peer Review in accordance to the classroom environment is only 37.5% applicable which is not suitable to deploy. Agent proposed Real Life Scenario would be a better strategy to deploy to improve poor engagement and emotional states of students in scenario D as the strength is 62.5% which is higher than Peer Review strategy strength followed by Online Quiz and Online Assessment which has the same strength as 50%.

TABLE VIII. RATING STUDENT BEHAVIOR FOR SCENARIO D

Engagement Factor	Indicators	Rating
Misbehavior	Verbally or physically aggressive	Medium
	Spoken in a raised voice	Medium
	Acting in an abusive manner	Not Present
	Yelling and screaming	Not Present
Poor motivation	Non-attentiveness	Medium
	Talking out of turns	Medium
	Playing video games	Medium
	Get away from class	Medium
Poor Participation	Avoiding eye contact	Medium
	Inability to initiate conversations	Medium
	Avoidance or refusal to participate	Medium

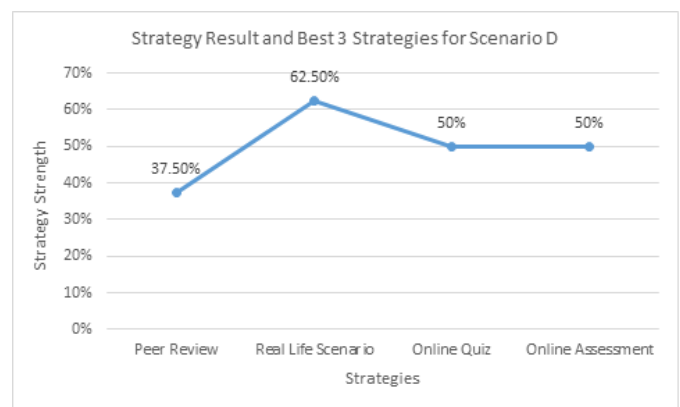


Fig. 12. Selected Strategy Result and Best 3 Strategies for Scenario D.

The Developed tool could provide the improvement to obtain teaching strategy easily compared to a traditional method by using software agents that automatically calculate the strength of each strategy according to environmental settings and also analyses the engagement level and emotional state of students. The values created by the tool include how strategy strength improves engagement and emotional state, reduces humans’ time, effort, and cost, and promising in the quality of teaching strategies. The result of the analysis from two different testing includes scenario-based testing and questionnaire feedbacks shows clearly how helpful the classroom simulator tool to the academician and social science researchers. Moreover, the results of the evaluation revealed that the respondents did not experience any difficulty while assessing the tool. The tool is easy to use and not complex as each function is well integrated.

IV. CONCLUSION LIMITATION AND FUTURE WORK

In this paper, we explored the use of agent-based social simulation and implemented the technology in evaluating the performance of students’ engagement in a classroom dynamics simulator. We looked at a typical traditional method used to study the poor student engagement issues for decades which consume considerable time, effort, and very limited settings. Furthermore, since these experiments are applied to humans’ society, there are very limited settings to be tested and the costs

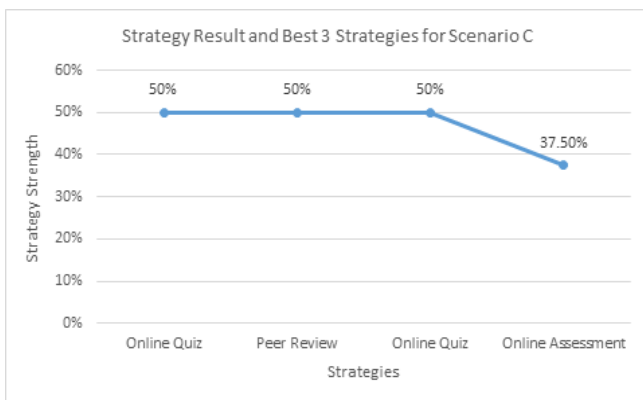


Fig. 11. Selected Strategy Result and Best 3 Strategies for Scenario C.

would be excessive if multiple settings are applied. The goal of this study is to automate the student engagement cycle using software agent technology to resolve some of the problems in the process. We demonstrated the use of software agents in measuring the engagement level and identifying the emotional states of students in a classroom dynamic simulator. On the other hand, the agent can calculate teaching strategy strength and recommends the best three strategies to academicians in accordance with their environmental factors.

Our contribution to this research is five-fold. Firstly, we identify three sets of attributes related to this study which are environmental attributes, engagement attributes, and emotional attributes that have been impacting teaching strategies. Then, we developed an agent-based evaluation model where we focus a great deal of our attention on emotional engagement and its associated attributes of existing strategies in enhancing poor students' engagement in classrooms. We thoroughly explored the theories and research findings on students' engagements area. Secondly, we propose a method to evaluate student engagement in a classroom dynamic simulator. We examine the three engagement factors that have been influencing engagement levels in a classroom such as misbehavior, poor motivation, and poor motivation. Thirdly, we examine the emotional factor; anger, anxiety, and boredom derived from engagement factors and study the issues and problems associated with them. The investigation covers the study and analysis of the work process including important aspects that lead our particular attention to which software agent technology could be applied. We propose a technique to identify emotional states that occurs among students' agents. Fourthly, we propose a technique to calculate a teaching strategy in accordance with environmental attributes includes the number of students, class duration and session, type of subjects and year of students, and emotional factors. On top of that, the agents can recommend the best three strategies to be deployed in certain environmental settings. Finally, we develop a solution in the form of a logical model for the classroom dynamic simulator in terms of the overall process flow, measurement of engagement level, analysis of emotional states of students, strategy selection and recommendation, and other novel aspects that could be deployed to improve the poor engagement performance in a classroom. In the proposed simulation, when the lecturer agent, detects low students' engagement, it selects a potential strategy to improve the low students' engagement. Therefore, the lecturer agent is aware of the attributes or specifications of every strategy defined in its knowledge base to decide on the potential one for a particular situation.

While the results of our work in the classroom dynamic simulator show considerable success in the research objectives, there are also some limitations and deficiencies in this work. However, these deficiencies do not compromise the significance of this research. The first limitation, the agent can produce strategy strength selected to particular environmental settings, however, we don't show the impact of strength or how the strategy strength can improve the engagement level due to lack of data. In this case, strategy strength and engagement level, and emotional states of students of the current situation stand as two different components. Another limitation, lecturer

emotion is equally important as students' emotions to complete the cycle of the process from the beginning of the selection of strategy, measuring engagement, identifying emotional states of the student, that eventually affect lecturer emotion which triggers them to revise strategy. However, our scope of the study is limited to examine students' emotional state, and we do not develop any formulation to examine or improve lecturer emotions. Another limitation, to determine the engagement level, we use the ordinal scale, nevertheless, a better and more realistic approach could be investigated here such as using fuzzy logic, which will be added to the limitation and future work of this study.

The scope for research in the classroom dynamic simulator overlays the way for many discoveries that could be integrated into agents. We outline here some interesting areas that could be investigated in our future work, (i) Virtual classroom designed with 3D graphical elements to portray emotions transition in colors and propose a technique to show how lecturer emotion being triggered and how it can be used to improve their emotions, (ii) Propose a method to show the interrelation between strategy strength and its ability to improve engagement level in a classroom, (iii) Use real data set from experienced academicians on proposing new strategies and show how the improvement of engagement level takes place.

ACKNOWLEDGMENT

This work is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No. J510050002.

REFERENCES

- [1] Owens, D. C., Herman, B. C., Oertli, R. T., Lannin, A. A., & Sadler, T. D. (2019). Secondary Science and Mathematics Teachers' Environmental Issues Engagement through Socioscientific Reasoning. *Eurasia Journal of Mathematics, Science and Technology Education*, 15(6), em1693.
- [2] Rissanen, A. (2018). Student engagement in large classroom: the effect on grades, attendance and student experiences in an undergraduate biology course. *Canadian Journal of Science, Mathematics and Technology Education*, 18(2), 136-153.
- [3] Inman, C. (2019). Examining Teacher-Student Relationships: Moving from Bullying to Caring (Doctoral dissertation, Southern Illinois University at Edwardsville).
- [4] Yazzie-Mintz, E. (2010). Charting the path from engagement to achievement: A report on the 2009 High School Survey of Student Engagement. Bloomington, IN: Center for Evaluation & Education Policy.
- [5] Bature, I. J. (2020). The Mathematics Teachers Shift from the Traditional Teacher-Centred Classroom to a More Constructivist Student-Centred Epistemology. *Open Access Library Journal*, 7(5), 1-26.
- [6] Thang, S. M. (2009). Investigating autonomy of Malaysian ESL learners: A comparison between public and private universities. 3L: Language, Linguistics and Literature, *The Southeast Asian Journal of English Language Studies*, 15, 97-124.
- [7] Cooper, K. M., Krieg, A., & Brownell, S. E. (2018). Who perceives they are smarter? Exploring the influence of student characteristics on student academic self-concept in physiology. *Advances in physiology education*, 42(2), 200-208.
- [8] Dasari, B. (2009). Hong Kong students' approaches to learning: cross-cultural comparisons. *US-China Education Review*, 6(12), 46-58.
- [9] Subramanian, L., & Mahmoud, M. A. (2020). A Systematic Review on Students' Engagement in Classroom: Indicators, Challenges and

- Computational Techniques. International Journal of Advanced Computer Science and Applications, 11(1), 105-115.
- [10] PISA, O. (2012). Results in Focus: What 15-year-olds know and what they can do with what they know. 2014-12-03]. <http://www.oecd.org/pisa/keyfindings/pisa-2012-results-overview>, pdf.
- [11] Yu, R., & Singh, K. (2018). Teacher support, instructional practices, student motivation, and mathematics achievement in high school. *The Journal of Educational Research*, 111(1), 81-94.
- [12] Pryor, J. H., Hurtado, S., DeAngelo, L. E., Blake, L. P., & Tran, S. (2010). *The American freshman: National norms fall 2009*. Univ of California Press.
- [13] Mann, S., & Robinson, A. (2009). Boredom in the lecture theatre: An investigation into the contributors, moderators, and outcomes of boredom among university students. *British Educational Research Journal*, 35(2), 243-258.
- [14] Nikish, C. (2013) Hobsons (<http://www.hobsons.com>) student engagement (<http://www.naviance.com/blog/c/studentengagementNT>).
- [15] Teoh, H. C., Abdullah, M. C., Roslan, S., & Daud, S. (2013). An investigation of student engagement in a Malaysian Public University. *Procedia-Social and Behavioral Sciences*, 90, 142-151.
- [16] Macal, C. M., & North, M. J. (2009, December). Agent-based modeling and simulation. In *Winter simulation conference* (pp. 86-98). Winter Simulation Conference.
- [17] Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental economics*, 21(1), 99-131.
- [18] Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: an international journal*, 46(3), 517-528.
- [19] Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90-102.
- [20] Pekrun, R., Cusack, A., Murayama, K., Elliot, A. J., & Thomas, K. (2014). The power of anticipated feedback: Effects on students' achievement goals and achievement emotions. *Learning and Instruction*, 29, 115-124.
- [21] Sharp, J. G., Sharp, J. C., & Young, E. (2020). Academic boredom, engagement and the achievement of undergraduate students at university: A review and synthesis of relevant literature. *Research Papers in Education*, 35(2), 144-184.
- [22] Lamborn, S., Newmann, F., & Wehlage, G. (1992). The significance and sources of student engagement. *Student engagement and achievement in American secondary schools*, 11-39.
- [23] Parsons, J., & Taylor, L. (2011). Improving student engagement. *Current issues in education*, 14(1).
- [24] Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement* (pp. 763-782). Springer US.
- [25] Hogan, R., Hogan, J., & Warrenfeltz, R. (2007). *The Hogan guide: Interpretation and use of Hogan inventories*. Hogan Assessment Systems.
- [26] Chin, K. O., Gan, K. S., Alfred, R., Anthony, P., & Lukose, D. (2014). Agent architecture: An overviews. *Transactions on science and technology*, 1(1), 18-35.
- [27] Jassim, O. A., Mahmoud, M. A., & Ahmad, M. S. (2015). A multi-agent framework for research supervision management. In *Distributed Computing and Artificial Intelligence*, 12th International Conference (pp. 129-136). Springer, Cham.
- [28] Mahmoud, M. A., & Ahmad, M. S. (2016, August). A prototype for context identification of scientific papers via agent-based text mining. In *2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 40-44). IEEE.
- [29] Mahmoud, M. A., & Ahmad, M. S. (2015, August). A self-adaptive customer-oriented framework for intelligent strategic marketing: A multi-agent system approach to website development for learning institutions. In *2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 1-5). IEEE.
- [30] Salleh, A. M., Desa, M. M., & Tuit, R. M. (2013). The Relationship between the Learning Ecology System and Students' Engagement: A Case Study in Selangor. *Asian Social Science*, 9(12), 110.
- [31] Subramainan, L., & Mahmoud, M. A. (2020, August). Academic Emotions Review: Types, Triggers, Reactions, and Computational Models. In *2020 8th International Conference on Information Technology and Multimedia (ICIMU)* (pp. 223-230). IEEE.
- [32] Subramainan, L., Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2017, June). A simulator's specifications for studying students' engagement in a classroom. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 206-214). Springer, Cham.
- [33] Subramainan, L., Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2016, August). A conceptual emotion-based model to improve students engagement in a classroom using agent-based social simulation. In *2016 4th International Conference on User Science and Engineering (i-USEr)* (pp. 149-154). IEEE.
- [34] Subramainan, L., Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2016, August). Evaluating students engagement in classrooms using agent-based social simulation. In *2016 2nd International Symposium on Agent, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 34-39). IEEE.
- [35] Subramainan, L., Mahmoud, M. A., Ahmad, M. S., & Yusoff, M. Z. M. (2016). An Emotion-based Model for Improving Students' Engagement using Agent-based Social Simulator. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6), 952-958.
- [36] Subramainan, L., Yusoff, M. Z. M., & Mahmoud, M. A. (2015, August). A classification of emotions study in software agent and robotics applications research. In *2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 41-46). IEEE.
- [37] Subramainan, L., Yusoff, M. Z. M., & Mahmoud, M. A. (2015, August). A classification of emotions study in software agent and robotics applications research. In *2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR)* (pp. 41-46). IEEE.

Wireless Body Area Sensor Networks for Wearable Health Monitoring: Technology Trends and Future Research Opportunities

Malek ALRASHIDI¹

Department of Computer Science, Community College
University of Tabuk, Tabuk, KSA

Nejah NASRI²

Laboratory of Electronics and Information Technology
(LETI), ENIS, Tunisia

Abstract—Today, there is an emerging interest in Wireless Body Area Sensor Networks (WBASNs) for the real-time monitoring of patients and early chronic disease detection. In this context, this paper presents a synopsis survey of healthcare monitoring via the IEEE 802.15.6 (UWB) protocol. We intend to propose a survey of the current issues of wearable physiological monitoring signals and devices, application areas, and reliability in WBASNs. To help elderly and disabled people, it would be beneficial to use a wireless transportable gadget at home to gather useful data in traditional human activities. This will manage regular hospital and emergency department appointments and will monitor crucial physiological signals real-time. This paper will also present a study on new wireless technologies intended for body area sensor networks, including signal processing problems, spectral allocation, security, and future research challenges of WBASNs.

Keywords—Healthcare; physiological signals; security; UWB; wireless technologies; WBASN

I. INTRODUCTION

Wireless Sensor Networks (WSNs) allow for better management of data collection in a healthcare context [1]. These include monitoring the patient's heart (electrocardiogram [ECG]), recording the electrical activity of the brain (electroencephalography [EEG]), evaluating the electrical activity produced by skeletal muscles (electromyogram [EMG]), and measuring the cornea-positive standing potential relative to the back of the eye (electrooculogram [EOG]). Physiological monitoring systems use an embedded WSN in the patient's biological tissue to regularly send the collected data to the base station, which performs the necessary processing to ensure the healthy functioning of vital organs [2]. These wireless technologies allow the person to permanently control these physiological parameters and their level of performance. These systems are used to send alarms to medical staff in the case of organ malfunction or abnormal parameter detection.

Existing signal control techniques used in hospitals cannot be used in transportable devices due to the following logic [3]:

- Regular systems for monitoring physiology are unwieldy and not healthy to wear for an extended period.

- There is loss of signal quality after extended use of the electrodes and gels.
- There are many cables used to transmit data to the sinks.

To resolve conflicts and difficulties in the field of controlling physiological signals, there is a need to develop wireless body area sensor network (WBASN) architecture, working on the invasive or non-invasive human body, to keep an eye on crucial health parameters.

A WBASN based on IEEE 802.15.6 standards [4] is a network composed of several vital signal sensing devices. The collected data from sensors are transmitted by wireless technologies to a medical server, then sophisticated applications analyze the data to make the right decisions based generally on artificial intelligence (machine learning, deep learning, and neuronal networks).

Technological advances in signal processing and the increasing level of integration for embedded systems have driven research into the engineering of new communication systems for the healthcare of elderly or disabled peoples, including WBASNs [5][6]. Figure 1 demonstrates the process of WBASN communication for health care monitoring.

Generally, communication in a WBASN goes through three steps [7]. The first step is the detection of vital organ parameters or physiological signals, such as blood oxygen level, blood pressure, ECG, EOG, and EEG. The second step is to transmit the collected data to the health server through a gateway using specific communication technologies, such as UWB, ZigBee, BLE, and Wi-Fi. The last step is responsible for classification, analysis, and early deduction of abnormalities. This step is achieved by using new artificial intelligence techniques, such as deep learning. Figure 2 shows a detailed description of the communication processes in WBASN.

In this paper, WBASN technologies are applied to critical healthcare monitoring research areas. In addition, the network architecture of WBASNs is presented. Resource allocation algorithms, routing, MAC protocols, security, and privacy are provided with qualitative comparisons.

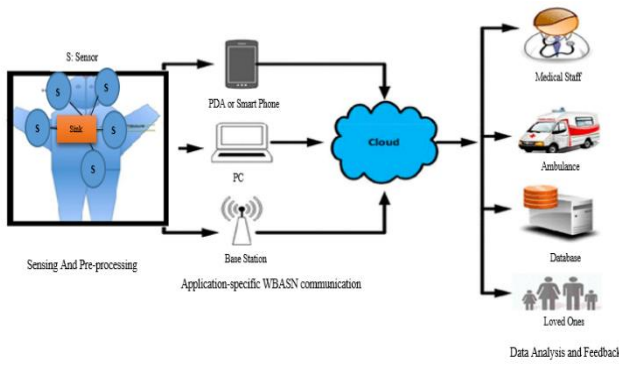


Fig. 1. Process of WBASN Communication.

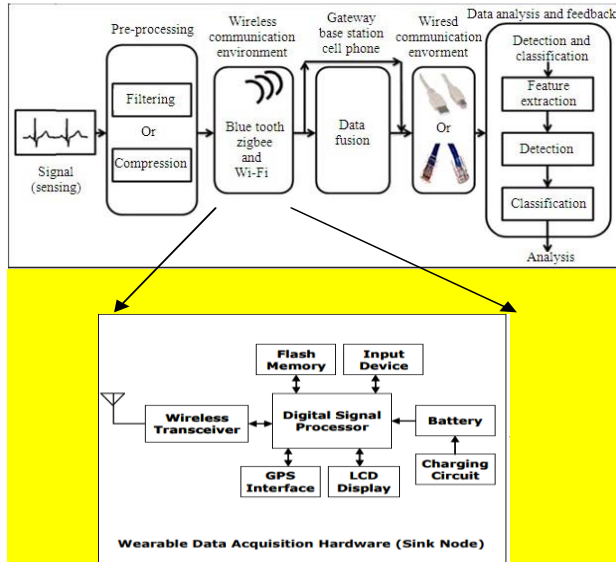


Fig. 2. Signal Processing and Communication in WBASN.

The rest of the paper is organized as follows. Section 2 introduces a synopsis of WBASN communication technologies. Section 3 explores WBASN architecture. Section 4 presents WBASN routing protocols. Finally, Section 5 challenges and explores issues in WBASNs.

II. WBASN COMMUNICATION TECHNOLOGIES

A WBASN is a sub-field of a WSN that interconnects sensor nodes or actuator capabilities in, on, or around a human body [8]. WPANs (wireless personal area network), like Bluetooth low energy (IEEE 802.15.1), Wi-Fi (based on 802.11b), UWB (Ultra-Wide Band) based on the 802.16.5 standard, and Zigbee (IEEE 802.15.4 Standard) can be used to ensure compatibility between nodes and the base station. All WBAN network architectures consist of several connected nodes, performing the tasks of communication detection and data processing.

A. Bluetooth Low Energy (BLE 4.0–5.0)

Bluetooth low energy is a specific technology characterized by low-power radio consumption in which applications using BLE can run on a small battery for many years. Despite traditional Bluetooth, BLE is designed to exchange data quickly and over long distances [9].

BLE operates in a 2.4 GHz frequency band and can transmit data over 40 channels using CDMA spread spectrum techniques. Additionally, BLE supports data throughput from 125 KB/s to 2 MB/S, and the power consumption level does not exceed 100 mw.

Table I describes the two versions of BLE, 4.0 vs. 5.0.

TABLE I. COMPARISON OF BLUETOOTH V 4.0 VS. V 5.0

	BLE 4.0	BLE 5.0
Data rates	125 Kb/s→1 Mb/s	125 Kb/s→2 Mb/s
Data ranges	10 meters indoors	40 meters indoors
Support IoT devices	No	Yes
Power consumption	High	Less
Security control	Less (16-bit CRC)	High (24-bit CRC)
Message length	≈255 bytes	≈31 bytes

B. ZigBee (IEEE 802.15.4)

ZigBee is a standard that needs a low-throughput, low-energy consumption, cost-effective wireless tenders, an extended lifetime battery, and trusted networking [10]. Technically, ZigBee technology is simpler than other common wireless standards, such as Bluetooth, Wi-Fi, and UWB. Comparing ZigBee with other technologies and standards, UWB is the most suitable for WSNs because of its low power consumption. ZigBee is proposed for short-range communication and limited energy consumption. Consequently, it will not affect the battery lifetime.

ZigBee presents two frequency bands, 2.4 GHz and 868/915 MHz. It gives 250 KB/s throughput data for 2.4 GHz; and 20 and 40 kbps for 868 and 915 MHz, respectively.

C. UWB (802.16.5)

Ultra-wide band (UWB) technology is characterized by low power consumption, high data throughput, and short-range communication [11]. Due to the health monitoring applications and limited coverage area, the IEEE 802.15.6 standard defines the physical and MAC layers of UWB.

Table II describes the characteristics of the IEEE 802.15.6 standard.

A summary statement of Bluetooth (IEEE 802.15.1), ZigBee (IEEE 802.15.4), and UWB (802.16.5) technologies is presented [9–12] in Table III.

TABLE II. CHARACTERISTICS OF THE IEEE 802.15.6 STANDARD

Data rate	Up to 10 Mb/s
MAC techniques	CSMA/CA Slotted Aloha
Communication range	≈10 meters
Security	Three levels of security: communication level, authentication level, encryption level
Frequency Band	402-405 MHz; 420-450 MHz; 863-870 Mhz;902-928 MHz; 950-958 Mhz;2360-2400 MHz; 2400-2485 MHz

TABLE III. TECHNICAL PARAMETERS OF ZIGBEE, BLUETOOTH (BLE.5.0), AND UWB TECHNOLOGIES

Standards	ZigBee	BLE 5.0	UWB
IEEE spec.	802.15.4	802.15.5	802.15.6
Frequency Band	2.4GHz	2.4 GHz	402 MHz→2485 MHz
Communication range	≈30 m indoors	≈40 m indoors	≈10 m indoors
Number of channels	27	40	27
CRC error detection	16-bit	24-bit	32-bit
Data rate	250 KB/s	Up to 2 MB/s	From 10 KB/s to 10 Mb/s.
Topology	star, tree, mesh	p2p, star	p2p, star

The UWB approach is intended for applications with reduced distances that are high throughput, such as health monitoring. In contrast, Wi-Fi is designed for high range and sustains devices with a significant power supply. The emerging technology of wireless sensor networks (WSN) has become a new paradigm for factory automation systems and industrial monitoring. The control of data size has been generally kept small (e.g., the temperature data in environmental control requests fewer than four bytes). Bluetooth and ZigBee technologies have provided excellent results, despite their slow data rate.

III. WBASN ARCHITECTURE

A WBASN is a new generation of WSNs adapted to support and control the human body. A WBAN consists of a certain number of nodes related to the physiological signals being monitored. Each node of the network is equipped with a biomedical sensor, processing unit (ADC, processor, etc.), and communication unit that generally use UWB (IEEE 802.15.6 standard).

Each network node, whether invasive or non-invasive, collects vital signals from the human body (ECG, EEG, etc.) and transmits them to a base station (central unit that processes the data using artificial intelligence techniques (deep learning, machine learning, neural network).

In addition to sensing and sending data, the body nodes can be equipped with actuators, such as pacemakers, to stimulate the heartbeat or a drug injection device.

A WBASN is an example of a WSN with variable characteristics, such as type of node, density of nodes, data transfer rate, mobility, interoperability, and communication distance. Moreover, the number of nodes in a body network depends essentially on the number of signals and organs to be controlled. Generally, this number is limited. Spies [12] presented the architecture of heartbeat control.

It is typically stable regarding the data transmission rates since it has real-time control of the physiological signals. They are associated with the patients and move together in the exact directions with the same mobility speed.

The architecture of a WBAN follows two models: a traditional model, if it is a single patient to be controlled, and a distributed model, if it controls several patients in a hospital.

For each model, there are three types of communication: intra-, inter-, and beyond-WBAN [13].

For each type of communication, there is a specific architecture. For example, for intro-WBAN communication, the adequate architecture is centralized (star topology).

For inter-WBAN communication, the most frequently used architecture is mesh architecture to guarantee low energy consumption and a better QoS. Finally, for beyond-WBAN communication, a mesh architecture is maintained to guarantee a better QoS.

A. Intra-WBAN Architecture

On-body communication among biomedical sensor nodes is called intra-WBAN communication. As shown in Figure 3, the intra-WBAN structure is made up of an invasive cluster head (CH) that collects data from invasive biomedical sensors in a star topology. CH relays information to the gateway and then to the medical server (MS) to process data. This architecture is done only for a single body [14].

B. Inter-WBAN and Beyond-WBAN Architecture

This architecture is designed especially for people or patients while exercising their quotidian routine, functioning indoors or outdoors.

Inter-WBAN architecture involves communication between two or more intra-WBANs [15] through a router or communication to transmit data to their destination, generally a medical server. In beyond-WBAN, architecture, authorized healthcare personnel (doctor, nursing orderly, etc.) have access to patients' medical information through the internet (cloud).

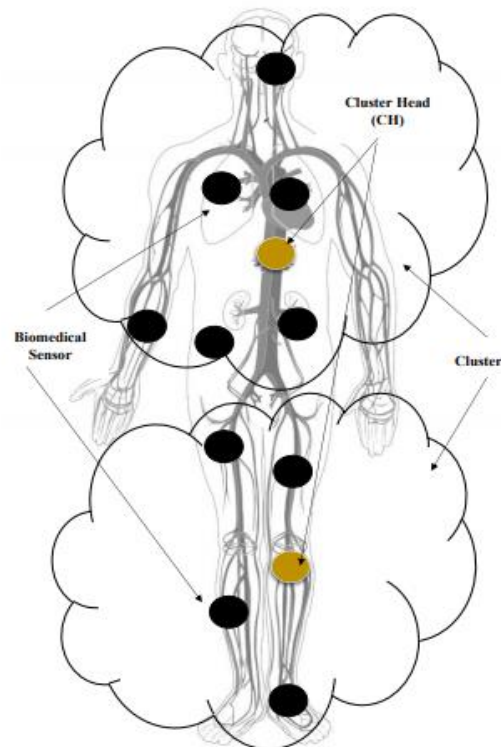


Fig. 3. Intra-WBAN Communication Architecture.

Figure 4 describes Inter/beyond-WBAN communication.

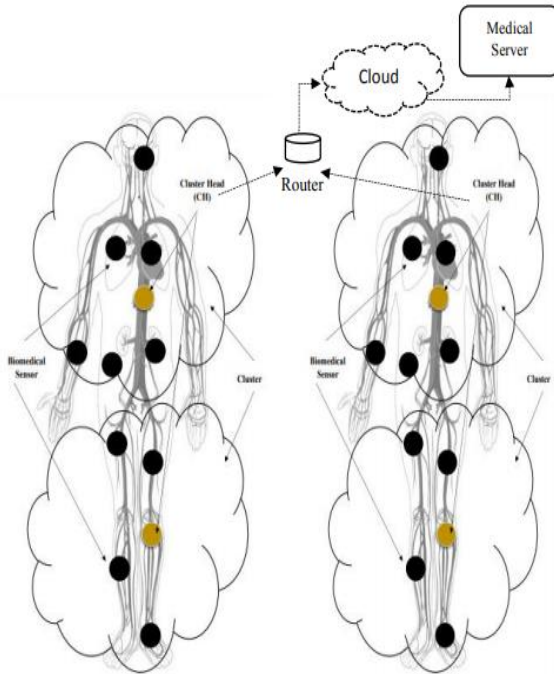


Fig. 4. Inter/Beyond-WBAN Communication Architecture.

IV. ROUTING IN WBASN

The operating environment of a WBASN is very complex and specific, especially for the transmission of physiologically relevant data that directly impacts the human being's life and health [16]. For these and other reasons, several challenges and constraints must be considered when designing a routing protocol to ensure the reliability of WBASNs. Among these constraints, the communication channel's specificity, whether invasive for an intra-WBASN architecture or non-invasive for inter and/or beyond architecture, is cited. This makes it necessary to consider biological tissue parameters for the frequency choice, allowing low transmission loss and reducing interference.

In addition, a WBASN is implanted in the human body. Therefore, recharging and replacing the batteries requires surgical intervention, and, in many cases, it is not possible to replace these batteries. For these reasons, battery life must be considered when designing the routing protocol. Similarly, the body nodes' temperature must be considered to protect the human body's biological tissues and organs.

The quality-of-service requirements must also be considered, especially since the sensors' data are of different types, such as necessity data, delay-sensitive data, and general data. According to recent research, WBASN routing protocols can be classified according to energy-, cluster-, temperature-, cross-layer-, QoS-, and posture-based routing.

A. Energy-based Routing

Minimizing energy consumption is a fundamental objective to be met when building an efficient WBAN architecture [17].

Indeed, proposing a routing protocol that regards the residual energy at the battery level allows.

Similarly, a routing protocol must consider the SAR (specific absorption rate) to increase the network's lifetime to considerably save human biological tissue. Most routing protocols in the literature only consider transmission energy, which represents two-thirds of the energy consumed, and do not consider the effects of electromagnetic waves absorbed by the human body (EMRF), which results in a high SAR. High energy consumption primarily affects the human body. It increases SAR, which affects biological tissue, reduces blood flow (circulation), and affects enzymatic reactions.

Power consumption has consequences on the WBAN network architecture in terms of lifetime and reliability, as well as on communication in terms of interference (EMI), propagation delay, reliability, and attenuation.

B. Cluster-based Routing

Self-organizations are the most studied problem by researchers [18]. One of them is based on the creation of a backbone to optimize the diffusion of information. A second solution called distributed hash table (DHT) aims to recover dynamic information [19].

Another approach called clustering techniques is based on dividing the network into small areas, called clusters. Each area is managed by a sensor node, called a cluster head (CH), with the aim being to optimize the network parameters. Several clustering algorithms have been defined. The election metric of CHs is not always the energy parameter, but node degree and mobility are also considered.

Much research has shown that the clustering techniques used for WSNs are well suited for WBASNs. Furthermore, many clustering techniques try to minimize direct communication between a biomedical node and the server to minimize the communication distance, subsequently reducing the communication energy dissipated by the transmitter.

C. Temperature-based Routing

An invasive biomedical node heats up during operation, which can damage vital human tissues and organs. Therefore, the temperature must be considered during the design of the routing protocols to avoid these problems through a node with a high temperature (Figure 5).

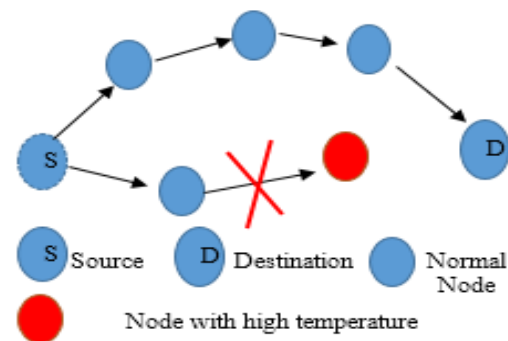


Fig. 5. Temperature-based Routing [20].

D. Cross Layer-based Routing

Cross-layer protocol refers to crossing or merging the functionalities of two or more layers in the protocol pile. The design of a routing protocol based on the cross-layer improves energy efficiency, interoperability between layers, QoS, and flow, according to the communication channel's characteristics. Typically, there are some protocols, such as thermal-aware routing algorithms and adaptive least temperature routing, named M-ATTEMPT [21].

The protocols are classified according to the layers of the protocol pile involved in the cross-layer strategy (physical and MAC layers; physical and network layers; physical, MAC, and network layers; physical link and network; physical link and application; physical network and application).

1) *Physical and MAC layers:* The objective of the interaction between these two layers is synchronization between the access techniques to the communication medium and the characteristics of the transceivers in terms of transmission power, SNR, and the energy level available to the batteries. For example, in the work presented in [22], energy-efficient cross-layer optimization was used for wireless personal area networks (WPAN), which were then deployed to WBANs, where the SNR value depended on the ARQ (automatic repeat request) of the MAC layer to increase energy efficiency.

2) *Physical and network layers:* SNR value, battery level, and electronic characteristics of the transceiver are considered when designing a routing protocol. It is an interaction between the physical layer and the network layer.

In [23], proposed a priority-based cross-layer routing protocol for healthcare uses. This protocol results from the interaction between the MAC layer and the network layer for invasive and non-invasive body communication. This work showed the reliability and stability of network reliability, throughput, QoS, and energy performance. In [25][26], protocols were proposed for these two layers to interact, improving network density and QoS.

E. Posture-based Routing

WBAN nodes are often disconnected from the network due to posture movement. Therefore, researchers regularly update a cost function that chooses the best path of packets to the sink. An example of a posture-based routing protocol is the on-body store and flood protocol, which guarantees better routing of information and low energy consumption based on multi-hop routing [24].

F. QoS-based Routing

The growing demand for real-time WBASN applications to support the elderly, disabled, and other patients is exponentially driving the need for QoS-based protocols. Moreover, the design of QoS-based routing protocols has a significant influence on the efficiency of WBASNs [25].

In a WBASN network, physiological signals have a direct impact on the health of the human being. For this reason, the QoS must consider the type of data to be transmitted.

Depending on the application, the routing process can integrate several QoS measures, such as propagation delay, energy consumption, and reliability.

In [26], a taxonomy of QoS-based routing protocols in WBASNs was presented. In addition, this work summarized the advantages and disadvantages of each routing protocol and presented a benchmark survey of all QoS-based routing protocols found in the literature.

V. OPEN ISSUES AND FUTURE RESEARCH

The increasing number of WBASN users and the growing number of wireless biomedical devices reveal several research topics and challenges for improving WBASNs. In this work, some open issues and future research directions, as described in [27][28][29], are presented.

A. Channel Allocation

The WBASN network's commercialization has increased the number of users, especially since the application areas are very diverse.

The network traffic in the communication channel increases with the number of users, which has augmented the collision rate, the loss of information, and the interference between the sub-channels for the different technologies used, such as UWB, ZigBee, and Bluetooth. The communication channel designed for WBAN may not meet future requirements.

A study by the Federal Communications Commission (FCC) has shown that specific frequency bands are partially occupied. Thus, cognitive radio (CR) can be a solution for the optimal utilization of communication channels.

RC is an intelligent communication technique that identifies and detects the used channel and moves communication to unused channels. CR optimizes the use of the signal spectrum to reduce interference between users.

The ISM band has become unable to support the requirements of WBASN, hence the idea of using unlicensed frequency bands outside the ISM band.

B. Network Security

A WBASN is a new technological trend allowing real-time control and monitoring of patients' health status. The security and protection of data against intrusions from detection to medical server routing is a crucial challenge for researchers in the field.

A WBASN architecture must use security measures that guarantee the confidentiality of medical records.

The main security and privacy challenges for a WBASN infrastructure are data authentication, confidentiality, integrity, freshness, availability of the network, secure management and localization, dependability, accountability, flexibility, privacy rules, and compliance requirements.

Security and confidentiality are the main challenges in WBASN architecture. Currently, several current research works focus on this problem, including privacy and quality of service (QoS), trust management, and integrating WBANs with mobile phones.

C. Radiofrequency Safety

For the internet of things, 5G refers to developments in terms of speed and services. The uses of 5G are diverse and varied, for example, the enrichment of the connected home, autonomous vehicle, and arrival of Health 4.0 (WBASN).

The commercialization of 5G has increased the number of electronic devices used in a body network. This has increased the rate of radiofrequency radiation (RF), causing electromagnetic interference (EMI), affecting sensor devices, and risking patient safety.

Despite standardization, the exposure limits to RF radiation must be reviewed because of the exponential increase of radiating devices with the commercialization of 5G and IoT service integration.

D. WBASN Service Standardization

WBAN technology commercialization requires interoperability between the different suppliers' electronic devices in the OSI model's layers.

Network scalability of WBAN devices should consider the maximum number of body sensors, energy consumption, packet losses, throughput, and synergic integration with 5G mobile communication.

E. Energy Consumption

WBASNs are used in the primary interest of human health (patient monitoring, disease detection, elderly monitoring, etc.). For this reason, reliability is a significant challenge. In contrast, energy consumption must be considered to guarantee the reliability of the network.

This parameter introduces the problem of energy conservation, especially when the application operates for a long time. Usually, it is not common for body sensor networks to recharge or replace node batteries after being depleted.

To increase the lifetime of a network or optimize the transmission parameters, research has proposed designing a virtual structure based on the phenomenon of self-organization. The main goal of self-organization is to minimize the transfer disorder in the network and improve the sensors' performance.

VI. CONCLUSION

In this study, a summary of the new collection of WSNs called WBASN is presented. There is a need for this technology for patient monitoring and maintenance in elderly and disabled people. Architecture routing protocols and a comparative survey between technology related to WBASN (ZigBee, BLE, UWB) are cited.

Additionally, problems related to the WBASN network and research directions to strengthen this technological trend for service integration with 5G mobile communication and a synergic integration with Industry 4.0 are presented.

REFERENCES

- [1] WU, Changcheng, et al. A low cost surface EMG sensor network for hand motion recognition. In: 2018 IEEE 1st International Conference on Micro/Nano Sensors for AI, Healthcare, and Robotics (NSENS). IEEE, 2018. p. 35-39.
- [2] OMODUNBI, Bolaji A., et al. Wireless sensor network based health monitoring system for hypertensive in-patients. *FUOYE J. Eng. Technol.*, 2018, 3.2: 6.
- [3] DARWISH, Ashraf; HASSANIEN, Aboul Ella. Wearable and implantable wireless sensor network solutions for healthcare monitoring. *Sensors*, 2011, 11.6: 5561-5595.
- [4] MARTELLI, Flavia; BURATTI, Chiara; VERDONE, Roberto. On the performance of an IEEE 802.15. 6 wireless body area network. In: 17th European Wireless 2011-Sustainable Wireless Technologies. VDE, 2011. p. 1-6.
- [5] NASRI, Nejah, et al. Efficient encoding and decoding schemes for wireless underwater communication systems. In: 2010 7th International Multi-Conference on Systems, Signals and Devices. IEEE, 2010. p. 1-6.
- [6] NICOLAU, Hugo; MONTAGUE, Kyle. Assistive technologies. In: *Web Accessibility*. Springer, London, 2019. p. 317-335.
- [7] NEGRA, Rim; JEMILI, Imen; BELGHITH, Abdelfettah. Wireless body area networks: Applications and technologies. *Procedia Computer Science*, 2016, 83: 1274-1281.
- [8] HÄMÄLÄINEN, Matti; LI, Xinrong. Recent advances in body area network technology and applications. *International Journal of Wireless Information Networks*, 2017, 24.2: 63-64.
- [9] FOURATI, Lamia Chaari; SAID, Sana. Remote Health Monitoring Systems Based on Bluetooth Low Energy (BLE) Communication Systems. In: *International Conference on Smart Homes and Health Telematics*. Springer, Cham, 2020. p. 41-54.
- [10] S. Mnasri, N. Nasri and T. Val, "An Overview of the deployment paradigms in the Wireless Sensor", *Networks International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN 2014)*, November 04–07th, 2014.
- [11] DE SANTIS, Valerio; FELIZIANI, Mauro; MARADEI, Francescaromana. Safety Assessment of UWB Radio Systems for Body Area Network by the $\{ \text{FD} \}^{\{ 2 \} \}$ Method. *IEEE Transactions on Magnetics*, 2010, 46.8: 3245-3248.
- [12] SPIES, Chel-Mari. Proposed model for evaluation of mHealth systems. In: 2015 International Conference on Computing, Communication and Security (ICCCS). IEEE, 2015. p. 1-8.
- [13] JABEEN, Tallat; ASHRAF, Humaira; ULLAH, Ata. A survey on healthcare data security in wireless body area networks. *Journal of Ambient Intelligence and Humanized Computing*, 2021, 1-14.
- [14] AL-JANABI, Samaher, et al. Survey of main challenges (security and privacy) in wireless body area networks for healthcare applications. *Egyptian Informatics Journal*, 2017, 18.2: 113-122.
- [15] ALI, Aftab; KHAN, Farrukh Aslam. Energy-efficient cluster-based security mechanism for intra-WBAN and inter-WBAN communications for healthcare applications. *EURASIP Journal on Wireless Communications and Networking*, 2013, 2013.1: 1-19.
- [16] UL HUQUE, Md Tanvir Ishtaique, et al. EAR-BAN: energy efficient adaptive routing in wireless body area networks. In: 2013, 7th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, 2013. p. 1-10.
- [17] SAGAR, Anil Kumar; SINGH, Shivangi; KUMAR, Avadhesh. Energy-aware WBAN for health monitoring using critical data routing (CDR). *Wireless Personal Communications*, 2020, 1-30.
- [18] MU, Jiasong, et al. A self-organized dynamic clustering method and its multiple access mechanism for multiple WBANs. *IEEE Internet of Things Journal*, 2018, 6.4: 6042-6051.
- [19] PUNJ, Roopali; KUMAR, Rakesh. CHS-GA: An approach for cluster head selection using genetic algorithm for WBANs. In: *Online Engineering & Internet of Things*. Springer, Cham, 2018. p. 28-35.
- [20] AHMED, Ghufuran; MAHMOOD, Danish; ISLAM, Saiful. Thermal and energy aware routing in wireless body area networks. *International Journal of Distributed Sensor Networks*, 2019, 15.6: 1550147719854974.
- [21] AHMAD, Ashfaq, et al. RE-ATTEMPT: a new energy-efficient routing protocol for wireless body area sensor networks. *International Journal of Distributed Sensor Networks*, 2014, 10.4: 464010.
- [22] CORREA-CHICA, Juan Camilo; BOTERO-VEGA, Juan Felipe; GAVIRIA-GÓMEZ, Natalia. Cross-layer designs for energy efficient

- wireless body area networks: a review. Revista Facultad de Ingeniería Universidad de Antioquia, 2016, 79: 98-117.
- [23] ELHADJ, Hadda Ben, et al. A priority based cross layer routing protocol for healthcare applications. Ad Hoc Networks, 2016, 42: 1-18.
- [24] NABI, Majid; GEILEN, Marc; BASTEN, Twan. MoBAN: A configurable mobility model for wireless body area networks. In: Proceedings of the 4th international ICST conference on simulation tools and techniques. 2011. p. 168-177.
- [25] KHAN, Zahoor A., et al. A QoS-aware routing protocol for reliability sensitive data in hospital body area networks. Procedia Computer Science, 2013, 19: 171-179.
- [26] QU, Yating, et al. A survey of routing protocols in WBAN for healthcare applications. Sensors, 2019, 19.7: 1638.
- [27] LIU, Qingling; MKONGWA, Kefa G.; ZHANG, Chaozhu. Performance issues in wireless body area networks for the healthcare application: a survey and future prospects. SN Applied Sciences, 2021, 3.2: 1-19.
- [28] CICIOĞLU, Murtaza; ÇALHAN, Ali. Channel aware wireless body area network with cognitive radio technology in disaster cases. International Journal of Communication Systems, 2020, 33.16: e4565.
- [29] OLATINWO, Damilola D.; ABU-MAHFOUZ, Adnan M.; HANCKE, Gerhard P. Towards achieving efficient MAC protocols for WBAN-enabled IoT technology: a review. EURASIP Journal on Wireless Communications and Networking, 2021, 2021.1: 1-47.

A-SA SOS: A Mobile- and IoT-based Pre-hospital Emergency Service for the Elderly and Village Health Volunteers

Kannikar Intawong¹, Waraporn Boonchieng²
Faculty of Public Health
Chiang Mai University
Chiang Mai, Thailand

Ekkarat Boonchieng⁴
Faculty of Science
Chiang Mai University
Chiang Mai, Thailand

Peerasak Lertrakarnnon³
Faculty of Medicine
Chiang Mai University
Chiang Mai, Thailand

Kitti Puritat⁵
Faculty of Humanities
Chiang Mai University
Chiang Mai, Thailand

Abstract—In Thailand, emergency illnesses are life-threatening conditions that constitute serious health problems and quick access to definitive care can improve the survival rate of the elderly dramatically. Currently, the pre-hospital emergency medical services have limitations which prevent the treatment from getting to the accident site on time. In this research, we proposed the A-SA SOS application, a mobile-and IoT-based pre-hospital emergency service for the elderly. The system helps the elderly to send the request to the nearest village health volunteers via a mobile application and smart device. After reaching the elderly, the village health volunteers help carry out basic life support to increase the survival rate before sending the patients directly to the Emergency Management System (EMS) agency. To evaluate the system, we tested it for three months in the Sub-district of Suthep in Chiang Mai city. Finally, the incident report showed that the average time to reach the scene (4.91 ± 0.56) to help elderly patients was less than the standard criteria of an average 3 minutes.

Keywords—Pre-hospital emergency service; mobile healthcare; IoT-based healthcare system; elderly; healthcare volunteer

I. INTRODUCTION

Currently, Thailand has become an aging society. According to the Institute for Population and Social Research, Mahidol University [1], in 2018, there were approximately 11.7 million Thai people aged over 60, accounting for 17.6 percent of the entire nation's population. It is extrapolated that the number of elderly will increase to 20 million in 2038. Emergency is a health problem that is common in elderly today. Apart from acute illnesses, chronic diseases, or geriatric syndromes, emergency illnesses can cause the elderly to become dependent or can lead to premature death. Based on statistical data reported during the year 2012-2015 of the Emergency Medical Institution of Thailand [1], the number of emergency illnesses increased annually from 1.1 million in 2012 to 1.3 million in 2015, 65 percent of all emergencies is caused by acute illnesses and 35 percent is caused by accidents. In addition, the deaths of the elderly who are

emergency patients outside the hospital tend to increase every year [2][3]. It was found that there were elderly deaths before receiving emergency services in 1,436 cases in 2013 and it increased to 1,786 cases in 2016. Although the number of elderly patients who died before the arrival of an ambulance squad to the scene (Response Time) within 8 minutes or more than 8 minutes after receiving an incident report from an elderly patient increases every year, the number of the deaths of critical elderly patients is more related to ambulance squads that spend more than 8 minutes to arrive at a scene than the number of deaths of such cases where the ambulance squad responds within 8 minutes. The development of an emergency medical system to be effective is essential to save lives and reduce the loss, including the disability of an emergency patient. The development of a system requires planning that is ready in all aspects of knowledge, personnel and equipment, including the technology to be used cost-effectively and efficiently. Previously, the National Reform Steering Committee on Public Health and Environment, and the National Reform Steering Assembly have established the Emergency Medical Institution as the main agency for establishing a single number emergency call center, and the Ministry of Education to put contents about first aid, basic resuscitation, and road safety in the curriculum of secondary education. The National Reform Steering Assembly has also created the Organization of Local Administration to operate and manage emergency medical services outside of hospitals in the local area [6]. Besides, an Out-of-hospital emergency operations guide has been developed for all levels to guide the practice of medicine that must be under the supervision of a medical professional staff and as a guide for medical practitioners to supervise emergency operators who are not medical professionals [4][5].

Therefore, a mobile system of volunteers for the pre-hospital emergency service is developed to assist medical emergencies for the elderly in urban communities through the association of the Organization of Local Administration and to

develop the capacity of emergency volunteers in the area to be able to provide pre-hospital treatment when an elderly person is in an emergency situation. That can reduce the effects of a severe illness that results in the elderly becoming dependent or in a subsequent death. In addition, we proposed prehospital emergency system called A-SA SOS which consisted of four systems: the A-SA SOS Rescuer application, the A-SA SOS elderly application, the A-SA SOS Smart device and the A-SA SOS pre-hospital data control center.

This research is organized as follows: Section I and II covers the introduction and the related works of the pre-hospital emergency system. In section III we propose the system architecture of the pre-hospital emergency system. Then, we evaluate the results of our proposed system in section IV. Finally, in section V, VI we summarize our research and discuss the future work.

II. RELATED WORKS

Mobile devices are present everywhere. Almost all of them have a computing power that is similar to a personal computer but their price is much lower. Their lightness is beneficial to various researchers for utilizing them to improve the effectiveness of rescue operations [7]. For instance, there is a mobile application created to aid PEMS officials to convey to patients with hearing loss [8]. They utilize mobile devices to develop the system form for testing the rescue operations' effectiveness [9]. There is an electronic note system developed for tablets used to record pre-hospital patient care [10].

The application called Emergency Medical Centre Locator (EMCL) is developed to assist patients in finding the nearest specialized emergency medical center. There are six specific sections that the application emphasizes, including injury, eye, cardiac, stroke, burn, and pediatric. This application is not available for Android, only available for iOS systems. Even though its name and description seem like it can help the patient find the nearest specialized emergency medical center, the application does not work based on this method. It allows the patient to find all the medical centers that appear and choose the place that is nearby and appropriate for the patient's emergency case. Actually, this application's function might not be good enough to be used in an emergency situation because of its inconvenience and because it takes too much time for a patient who requires instant help. Nonetheless, there is no genuine emergency help given as this application does not offer any facilitations to a patient emergency call for rescue and is unable to convey to physician [13].

A model has been developed with the purpose of supporting senior citizens' healthcare and handicapped people. The model is grounded on personal-centric sensing structure. It offers the elderly and handicapped people with the service that serves for emergency responsiveness when there is an uncommon health condition [12].

Pre-hospital Emergency Notification System has been developed for mobile platforms to enable the emergency medical officials to notify a team in the hospital about

individual information of the incoming casualty and about the seriousness of injuries of road accidents. Moreover, the system is able to let the hospital officials learn about the information of the incoming injuries. Pre-hospital Emergency Notification System has been created for a mobile application, using 2 inventor apps, such as MYSQL and PHP [14].

There is a system created to assist the elderly to locate any nearby medical place simply by using the information of a mobile GPS locality along with access point. Furthermore, they can observe users with a Bluetooth beacon. Hence, the system will automatically evaluate users' present coordinates with the GPS or the information of a network system. This means that the system can locate any nearby medical centers or clinics from the database that stores the information about the medical places' location [11].

III. THE PROPOSED SYSTEM

A. A-SA SOS Users

The A-SA SOS users are categorized into four types of participants: elderly with smartphone, elderly with SOS smart device, Village Health Volunteers (VHV), A-SA SOS officer. In addition, the term "elderly with smartphone" refers to elderly people who are familiar with using smartphones and need medical emergency request for the rescue. "Participants with SOS smart devices" refers to the elderly who are using the SOS smart device instead of smartphone. "Village Health Volunteers" (VHV) is a participating group who takes care of the health of people in rural areas. A-SA SOS officer is the person who manages the overall functioning of the system for the elderly and the Village Health Volunteers.

B. System Architecture

The A-SA SOS application aims to improve the pre-hospital emergency service for people living in urban cities. The system architecture of A-SA SOS is shown in Fig 5. As we mentioned before, there are four types of actors in the system. For each actor, we proposed three system architectures in order to improve the pre-hospital emergency service, namely, A-SA SOS Rescuer application, A-SA SOS elderly application and A-SA SOS pre-hospital data control center. The details of each system can be described as follows:

1) *A-SA SOS Rescuer* application is focused on the Village Health Volunteer user around rural areas near the elderly people and it performs several functions as follows: First, it receives the emergency rescue request from the A-SA SOS elderly application via a cloud server then it uses the elderly geo location to reach for a first aid and rescue operation. Second, it enables to contact and point out the elderly location and forwards it to medical equipment transporters near the emergency site in serious cases. Third, the application reports on the progress of the rescue situation and transfers the elderly information to the A-SA SOS pre-hospital center then the system contacts the hospital which is suitable for the elderly. The A-SA SOS Rescuer application is shown in fig 1.

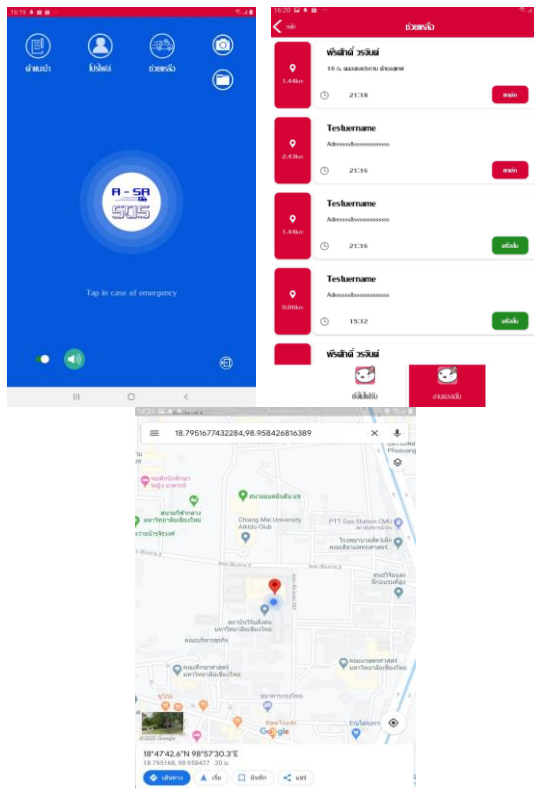


Fig. 1. A-SA SOS Rescuer Application.

2) A-SA SOS elderly application is proposed to mainly focus on the elderly of traumatic and non-traumatic cases who live at home alone or are bedridden. The application consists of two types of rescuer platform.

a) *A-SA SOS mobile application*: This platform is designed for elderly who are familiar with using smartphone applications. The application was developed to send the rescuer request together with the geoinformation of the user, and it is designed to have only one button on the user interface. The application was developed by the Xamarin cross platform version 4.2 and it sends data to the cloud server by JSON data format via https protocol. The A-SA sos application is shown in Fig. 2.

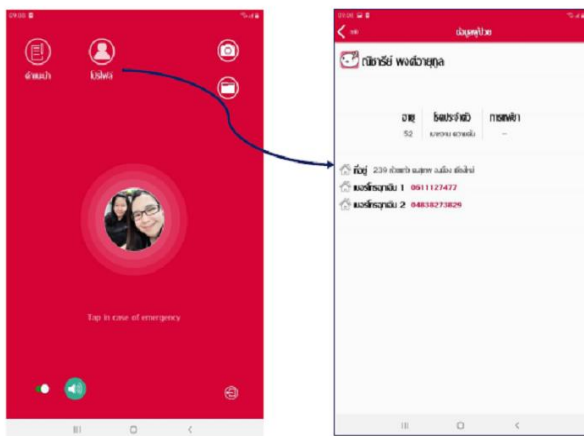


Fig. 2. A-SA SOS Mobile Application on Android os for Patients.

b) *A-SA SOS Smart device*: The IoT-based platform for smart device is designed for the elderly who don't use a smartphone. For this case, we developed a smart device based on an arduino with 4G sim card which sends data to the cloud server by MQTT protocol. The device is designed with one button on the center of the device in order to be usable in case of an accident. The A-SA SOS smart device is shown in fig 3.

c) *A-SA SOS officers and the Resuscitation team*: This web-based platform is to support the back-office for the A-SA SOS officers. In terms of the management system, we designed a system based on corporate management among elderly people, Village Health Volunteers and the Resuscitation team in order to manage the emergency cases of the elderly until it is confirmed that they accessed pre-hospital service.



Fig. 3. A-SA SOS Smart Device.

d) *A-SA SOS pre-hospital data control center*: The core system of the pre-hospital support center was developed based on the cloud server consisting of three modules: First, the Broadcasting module is responsible for broadcasting the elderly information while they are in need of being rescued by Village Health Volunteers. Second, the Web-based Monitor Module is the core data center that computes the possible matches between the elderly and Village Health Volunteers and synchronizes it with the database system. Finally the A-SA SOS Management Module is responsible for querying information of all users in the system from the database. The overall system architecture of A-SA SOS is shown in Fig 4.

C. Procedure Scenario Example

In this section, we show the scenario of the procedure of how the A-SA SOS system works. The procedure consists of the steps to be followed if an accident or sudden illness occurs

- 1) The elderly pushes the rescue button of the application or the button of the smart device in case the user is not familiar with the smartphone, in order to call for rescue.
- 2) The data from the application or device used by elderly is sent to the cloud server then the Web-based Monitor Module is computed for the top ten list of Village Health Volunteers who are located near to the elderly.
- 3) The Web-based Monitor Module computes the distance between the geo-location of the elderly and that of all the Village Health Volunteers by the Haversine formula[15]. The equation (1) is described below.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\theta_2 - \theta_1}{2}\right) + \cos(\theta_2) \cos(\theta_1) \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad (1)$$

where

ϕ_1, ϕ_2 are the latitude of point 1 and latitude of point 2 (in radians)

λ_1, λ_2 are the longitude of point 1 and longitude of point 2 (in radians)

We filter the list of Village Health Volunteers to those not further than 1.5 kilometres and create a priority list of not more than ten volunteers. Then, the broadcast module sends the emergency case to the A-SA SOS Rescuer application.

4) In this step, the Village Health Volunteer who uses the A-SA SOS Rescuer application receives a notification of the geo-information of the elderly. Then a matching between a Village Health Volunteer and the elderly is made depending on who pushed on the system first.

Please note that, in case the system is unable to make a matching to rescue the elderly within three minute two times in row then the resuscitation team or 1669 is responsible for the case.

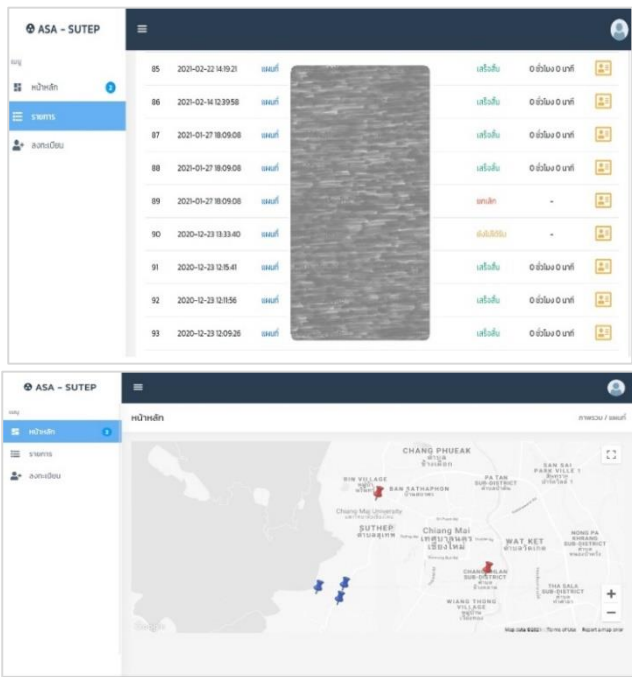


Fig. 4. A-SA SOS Officers and Resuscitation Team.

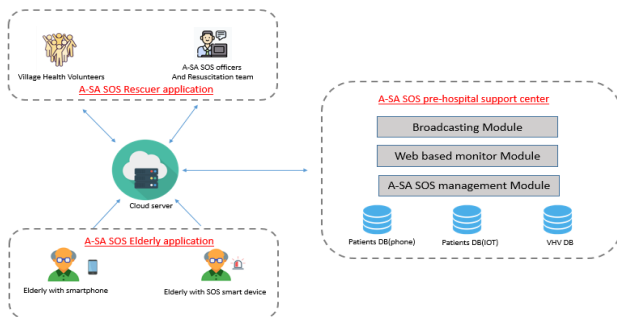


Fig. 5. A-SA SOS System Architecture.

5) Next, the Village Health Volunteer reaches the elderly and gives first aid. Then symptoms are assessed and forwarded to the resuscitation team or 1669 in serious cases

From the procedure scenario f example, it can be seen how the elderly can benefit from our approach by using the A-SA SOS system. When an elderly person gets in an accident she can immediately contact the nearest village Health Volunteers and the resuscitation team or 1669. Also, officers can get all the information needed to rescue the elderly from the application. Most importantly, all these procedures can drastically reduce the operation time and greatly increases the survival rate of the elderly.

IV. EXPERIMENT AND RESULTS

This section shows the evaluation of our experiment and the result of the proposed A-SA SOS system. To evaluate the outcome, we carried out a pilot test of the A-SA SOS system in real situation for three months in the Sub-district of Suthep in Chiang Mai city. There were 236 participants totally, with 171 female (74.34%) and 59 male (25.65%) participants. From all the participants, there were 30 elderly who used the A-SA SOS Smart device. The statistics of the total number of time for using the emergency of our proposed system is shown in table 1.

From the usage of a medical emergency rescue system for the elderly in Suthep, Mueang, Chiang Mai, between August 1 - October 31, 2020, it was found that the emergency operations of Suthep Sub-District Municipality totally sent a medical emergency rescue for 226 times. Most of service users were under aged 60, total of 170 times, 45 times were elderly patients over 60, and 11 times were unable to specify their age. For the elderly, 62.22% were female and 77.78% lived in Suthep Sub-District Municipality. There were five causes of emergency medical services, including 1) illness, fatigue (non-specific) 22.22%, 2) Dyspnea 17.78%, 3) fall, accident, pain 15.56%, 4) cardiac arrest 11.11% and 5) unconsciousness, unresponsiveness, faint 11.11%. For the duration of the operation, it was found that the incident report time to the order time took 0.31 ± 0.09 minutes, the incident report time to departure time took 4.60 ± 0.64 minutes, and the incident report time to a scene took 4.91 ± 0.56 Minutes (Table 1).

TABLE I. THE TOTAL USAGE TIME OF THE EMERGENCY OF OUR PROPOSED SYSTEM

Information	Age > 60 years		Age < 60 years	
	Number	Percentage	Number	Percentage
Gender				
Male	17	37.78	87	51.18
Female	28	62.22	83	48.82
Symptoms				
stomachache, backache, pelvis, and groin pain	1	2.22	10	5.88
Anaphylaxis and allergic reaction	0	0	7	4.12
animal bite	0	0	0	0
bleeding (without injury)	0	0	2	1.18
Dyspnea	8	17.78	5	2.94

Cardiac arrest	5	11.11	2	1.18
Angina Pectoris	0	0	2	1.18
choking	0	0	0	0
Diabetes	1	2.22	1	0.59
environmental health hazards	0	0	0	0
wounded	1	2.22	1	0.59
headache, sore throat	0	0	1	0.59
Manic episode, neurosis, temper	1	2.22	11	6.47
overdose, poisoning	0	0	7	4.12
pregnancy, delivery, gynecology	0	0	0	0
seizures	0	0	2	1.18
illness, fatigue (non-specific) etc.	10	22.22	18	10.59
limb weakness, Dysarthria, crooked- mouth	4	8.89	5	2.94
unconsciousness, unresponsiveness, faint	5	11.11	7	4.12
child, infant (Pediatrics)	0	0	0	0
injured	1	2.22	24	14.12
burn (hotness), electricity, chemical	0	0	0	0
drowning, diving injury, water injury accidents	0	0	0	0
fall, accident, pain	7	15.56	37	21.76
Motor vehicle accident injury	1	2.22	28	16.47
Operating area				
In Suthep Sub-district Municipality	35	77.78	134	78.82
Outside Suthep Sub-district Municipality	10	22.22	36	21.18
Time of operation				
Incident report order time (minute)	0.31±0.09		0.24±0.04	
Incident report departure time (minute)	4.60±0.54		5.22±0.59	
Incident report time to arrive at scene (minute)	4.91±0.56		5.46±0.56	

V. DISCUSSION

From the results of this study, it can be concluded that the application of the medical emergency rescue system for the elderly in urban communities has been used to help the elderly

report emergencies effectively and quickly and emergency rescuers could also reach the scenes faster. The medical emergency rescue system includes 1) A-SA SOS elderly application and smart devices were developed for the elderly to report emergency incidents. After reporting an emergency incident, the device sends signals via cellular wireless communication and notifies the location, and sends information to the emergency volunteer application (A-SA SOS Rescuer application) and the emergency volunteer center (A-SA SOS officers and Resuscitation team:). According to the device usage of the elderly for 3 months, it was found that the developed devices may be difficult to use for the elderly in terms of convenience, as the batteries last for approximately 15 hours, which must be recharged every day before the battery runs out. If the battery is completely discharged, the device must be turned on again. 2) The emergency volunteer application (A-SA SOS Rescuer application) is an application used for reporting and receiving emergency reports for general citizens and community emergency volunteers. This application is available only on the Android system, not on the IOS system. When evaluating the possibility of using the app, it can notify the incident time more quickly because the app will send signals to the application of a person directly and can tell the location of the emergency scene. 3) From the rehearsal of the aid system that has been developed, it was found that the incident report time to the order time was less than 1 minute, which was compatible with the standards of the National Institute of Emergency Medicine. Finally, the reported average time to arrive at a scene (4.91 ± 0.56) when helping elderly patients was less than the standard specified criteria of 8 minutes with an average of 3 minutes.

VI. CONCLUSION AND FUTURE RESEARCH

In this research, we proposed the A-SA SOS system which consisted of a mobile and IoT-based healthcare application for the pre-hospital medical emergency rescue system for the elderly in urban communities. The main contribution of this research is to provide an effective management tool for the Village Health Volunteers to accelerate and improve the reach of elderly in order to give faster first aid during an emergency situation. In terms of evaluation of the system, we tested the A-SA SOS system for three months in the Sub-district of Sutep in Chiang Mai city. There were 226 medical emergency rescue cases in total and according to the incident report the average time to arrive at a scene was (4.91 ± 0.56) when helping elderly patients which was 3 minutes less in average than the standard criteria.

In the future work, we first plan to improve the A-SA SOS Smart device in terms of convenience such as usability and extended battery lifetime. Second, we would like to test the system in the wider area of Chiang Mai and for a longer period than 3 months. Third, we plan to improve the application A-SA SOS Rescuer application to develop the automatic rescue request notification using a messaging platform API such as LINE, Facebook to provide health volunteers with more convenience when receiving elderly calls in emergency situations.

ACKNOWLEDGMENT

This work was supported by the National Research Council of Thailand (NRCT).

REFERENCES

- [1] Prasartkul, P., Rittirong, J., Chuanwan, S., Kanchanachitra, M., Katewongsa, P., & Thianlai, K. (2015). Situation of the Thai elderly. Nakhon Pathom: Institute for Population and Social Research Mahdiol University.
- [2] Sukkird, V., & Shirahada, K. (2018). E-health service model for Asian developing countries: A case of emergency medical service for elderly people in Thailand. In *Optimizing Current Practices in E-Services and Mobile Applications* (pp. 214-232). IGI Global.
- [3] Shatpattananunt, B., Wiangosot, S., & Pintatham, K. (2018). Elderly patients' experience in using emergency medical service at a tertiary hospital, upper northern Thailand. *Songklanagarind Journal of Nursing*, 38(3), 102-115.
- [4] Smith, C. M., Wilson, M. H., Ghorbangholi, A., Hartley-Sharpe, C., Gwinnutt, C., Dicker, B., & Perkins, G. D. (2017). The use of trained volunteers in the response to out-of-hospital cardiac arrest—the GoodSAM experience. *Resuscitation*, 121, 123-126.
- [5] Vattanavanit, V., Uppanisakorn, S., & Nilmoje, T. (2020). Post out-of-hospital cardiac arrest care in a tertiary care center in southern Thailand: from emergency department to intensive care unit. *Hong Kong Journal of Emergency Medicine*, 27(3), 155-161.
- [6] Pudpong, N., Durier, N., Julchoo, S., Sainam, P., Kuttiparambil, B., & Suphanchaimat, R. (2019). Assessment of a voluntary non-profit health insurance scheme for migrants along the Thai–Myanmar border: a case study of the migrant fund in Thailand. *International journal of environmental research and public health*, 16(14), 2581.
- [7] Toahchoodee, M. (2017, July). ARSA-the pervasive Rescuer Supporting System for the Pre-Hospital Emergency Medical Service. In 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.
- [8] F. Buttussi, L. Chittaro, E. Carchietti, and M. Coppo, "Using mobile devices to support communication between emergency medical responders and deaf people," in *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '10. New York, NY, USA: ACM, 2010, pp. 7–16. [Online]. Available: <http://doi.acm.org/10.1145/1851600.1851605>
- [9] K. Nakata, K. Maeda, T. Umedu, A. Hiromori, H. Yamaguchi, and T. Hi-gashino, "Modeling and evaluation of rescue operations using mobile communication devices," in *2009 ACM/IEEE/SCS 23rd Workshop on Principles of Advanced and Distributed Simulation*, June 2009, pp. 64–71.
- [10] A. F. Duarte, H. V. Cesar, A. L. M. Marques, P. M. d. A. Marques, and G. A. P. Junior, "Prehospital electronic record with use of mobile devices in the samu's ambulances in ribeirão preto–brazil," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, June 2015, pp. 362–363.
- [11] Huh JH, Kim TJ (2018) A location-based mobile health care facility search system for senior citizens. *The Journal of Supercomputing*
- [12] Hussain A, Wenbi R, da Silva AL, Nadher M, Mudhish M (2015) Health and emergency-care platform for the elderly and disabled people in the Smart City. *Journal of Systems and Software*
- [13] Moskowitz (2014) LLC emergency medical center locator. Available at: www.itunes.apple.com
- [14] A. Sarlan, F. K. Xiong, R. Ahmad, W. F. W. Ahmad and E. Bhattacharyya, "Pre-hospital emergency notification system," *2015 International Symposium on Mathematical Sciences and Computing Research (iSMSC)*, Ipon, 2015, pp. 168-173.
- [15] Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly*, 64(1), 38-40.

An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks

P. Bharath Kumar Chowdary^{1*}

Research Scholar, Department of Computer Science and Engineering, BIST, Bharath Institute of Higher Education and Research (BIHER), India

Dr. R. Udaya Kumar²

Research Supervisor, Professor, Department of Information Technology, BIST, Bharath Institute of Higher Education and Research (BIHER) Institution, India

Abstract—The most common disorder affecting millions of population worldwide due to insufficient release of insulin by pancreas is diabetes. Early detection or precaution of diabetes is necessary, otherwise leads to many complicated problems. Predicting diabetes at early stages with appropriate treatment, individuals can maintain a happy life. If the conventional diabetes detection method is tedious, the identification of diabetes from clinical and physical data requires an automated system. This paper proposes an approach to enhance diabetes prediction using deep learning techniques. Based on the Convolutional Long Short-term Memory (CLSTM), we developed a diabetes classification model and compared with the existing methods on the Pima Indians Diabetes Database (PIDD). We assessed the findings of various classification approaches in this study. The proposed approach is further improved by an efficient pre-processing mechanism called multivariate imputation by chained equations. The outcomes are promising compared to existing machine learning approaches and other research models.

Keywords—Convolutional long short-term memory; diabetes prediction; machine learning; pre-processing

I. INTRODUCTION

Diabetes is affecting the world's elderly population in a very drastic way [1]. By 2019, 463 million individuals around the globe had diabetes. It is expected by the International Diabetes Federation (IDF) that the number of patients rises to 700 million individuals in near future.

Diabetes occurs due to the inconsistency of glucose levels in the blood. Usually, diabetes is classified into type 1 and type 2 diabetes. Type 1 diabetes is due to little insulin production and type 2 occurs due to blood cells becoming insulin resistant. The fundamental cause of diabetes remains unclear, but scientists agree that diabetes plays a significant role in both genetic factors and environmental lifestyles. And though it is incurable, therapy and medicine can handle it by maintaining the levels in check.

Diabetes slowly causes different diseases in the long run. Mainly it affects the heart, nervous systems, retina, kidneys and other internal organs. The care taken at the early stages of diabetes helps in avoiding the damage of various organs. Although it is a chronic problem, researchers handled this by developing various prediction systems using machine learning

algorithms [2]. The most popular algorithms were Support Vector Machine (SVM), Decision Trees, and Random Forest.

Another popular model to predict diabetes is an Artificial Neural Network (ANN) [2]. It is well-known for its high precision and performance. Present research includes Deep Learning (DL) for prediction due to the increasing size and complexity of data.

Recent studies [3] using DL have enhanced various prediction and classification parameters like accuracy and precision. PIMA diabetes dataset [4] is used by many researchers to test their models.

Diabetes occurs when the body is unable to metabolize the glucose. The body is unable to produce or react to the insulin produced in the case of diabetes. Once diabetes is attacked, it is tough to cure. Hence, the knowledge of how diabetes occurs helps individuals to prevent it. Early diagnosis helps in reducing the risk for the patient.

Practitioners require high amount of data. The healthcare industry collects a large amount of health-related data, but this data cannot perceive undetected patterns of good decision-making [5]. It is a tedious job for any individual to process a high amount of data. As a result of this, researchers developed various machine learning and classification techniques to handle the data.

This paper has used Traditional LSTM and convolutional LSTM models for prediction on the PIMA dataset. We have performed extensive experimentation using data mining algorithms such as decision trees (DT), Naïve Bayes classification, ANN, and DL to provide an insight into how different algorithms work for diabetes prediction. In a logical and well-organized way, the comparison of algorithms is interpreted, with more efficient and prominent results provided by DL. DL is a self-learning framework for knowledge used successfully to predict diabetes.

II. RELATED WORK

A. Diabetes Prediction using Machine Learning (ML) Algorithms

ML algorithms are used by researchers to predict diabetes. The most famous approaches are SVM, J48, K-Nearest Neighbours (KNN), and Random Forest classifiers [8]. Ioannis

*Corresponding Author

et al. [7] applied ML and data mining (DM) techniques for diabetes prediction. This work [7] mainly focused on analysing the existing techniques in ML and DM. The authors have done extensive research on different databases containing diabetic data.

Zhu et al. [9] used a logistic regression-based model to predict diabetes. The authors have used principal component analysis and k-means algorithms to classify the developed model data correctly. The authors [10] developed a prediction model on diabetes data using classifiers based on the decision tree, naïve Bayes and random forest.

The authors [11] also used various MLK algorithms to classify diabetes data. This work [11] majorly focused on using decision trees and SVM to classify PIMA diabetes data. Dataset partitioning is carried out using a 10-fold method of cross-validation. The authors have not performed data pre-processing.

Negi and Jaiswal [12] also applied SVM to diabetes prediction on PIMA and Diabetes 130-US datasets.

The authors tested the existing ML algorithms on various datasets to predict diabetes. But the data consists of missing values and requires pre-processing. We are using data pre-processing techniques to enhance diabetes classification. The next part of this section covers various deep neural network models for diabetes prediction.

B. Deep Neural Networks

In the analysis of large datasets, researchers have begun to realize the capabilities of DL techniques [6]. Therefore, using DL techniques, diabetes prediction has also been carried out.

The authors [13] used a Deep Neural Network (DNN) for diabetes prediction. This approach was tested on the PIMA dataset. As DNN can filter the data and develop biases, the authors did not deliberately pre-process the dataset. For the research collection and the rest of the research, the dataset is divided into 192 samples. 88.41 percent was the accuracy rate stated by the authors.

Another approach [14] based on CNN and CNN-LSTM is developed to test the Electrocardiograms dataset.

The authors [15] used the logistic regression model as a basis for the multilayer neural network and CNN. The dataset used by authors [15] consists of nine patients. For each patient nine features are gathered. Moreover, each patient had data for 10,800 days, resulting in a total of 97,200 simulated days. There was no proper discussion of the attributes used in this analysis.

Miotto et al. [16] proposed the Deep Patient model, which is an unsupervised DNN. This model is used to classify electronic health records. The model is tested on a database consisting of 704,857 patients.

The authors [17] tested various deep learning methods on Australian hospital health records and developed a dataset.

The authors [18] used RNN model to predict both type 1 and type 2 diabetes. The authors used the PIMA dataset and predicted that the attribute "Glucose" has the highest

significance followed by BMI, age, births, pedigree feature of diabetes, blood pressure, thickness of the skin and insulin.' The training dataset and 20 percent for the testing were split into 80 percent to validate the analysis.

This paper proposes a convolutional neural network with enhanced feature selection and data pre-processing mechanisms for diabetes prediction. The later section provides the proposed methodology.

The existing models fail to extract the features properly. The existing models failed to properly incorporate the data pre-processing techniques. The existing models do not fill the missing values. Moreover, neural networks and error propagation are not implemented by existing models. The proposed model overcomes all the above-mentioned problems and enhances the Diabetes prediction task. The remaining part of the paper is as follows. Section 2 gives the proposed methodology. The fourth section gives dataset description and selected results of the existing and proposed method.

III. MATERIALS AND METHODS

90 percent of all forms of diabetes are diabetes types II. This disorder causes insulin resistance or insulin loss problems for the victim. The age at which diabetes type II typically takes place is 40 years old. Youth under the age of 30 are at risk for this disease with current eating habits and lifestyle. Early detection with routine checks and surveys allows people to diagnose the disease early and to take precautions.

Various research attempts were made to enhance the accuracy and applicability of various Clinical Decision Support Systems (CDSS) interpretability. However, it is still essential to optimize this issue. In the medical area, where interpretability is an essential question, fluid rules are relevant.

Many healthcare systems gain valuable information and produce a huge amount of clinical data. Machine learning techniques allow the practitioner to process this data and make quick decisions [9]. These decisions reduce the risk of diabetes, affecting the person severely, and preventing damage to other organs. Multiple machine training techniques for disease prediction and information from medical data have been developed.

The long short-term memory (LSTM) [21] is a form of RNN and consists of feedback connections. LSTM models can process a long input data sequence at ease.

A standard LSTM system consists of a cell, an entrance gate, an output gate, and a forgotten gate. The cell recalls values at arbitrary times, and the three gates monitor information flow in and out of the cell.

LSTM networks are well suited for the classification, processing, and estimation of time series data because the period of uncertain events in a time series can be delayed. LSTMs have been developed to resolve the disappearance gradient problem that can be observed during conventional RNN training. Relative lack of attention to the length of gaps is an advantage of LSTM in multiple applications over RNNs, hidden Markov models, and other sequence learning methods.

Compared with a popular recurrent unit, an LSTM cell has the benefit of its cell memory unit. The cell vector can encapsulate the concept of missing some of its formerly saved memory and add some of the new details. The cell equations and the sorting of sequences under the hood must be inspected to demonstrate this.

A. Traditional LSTM

A LSTM network comprises of memory cell and four gates. The four gates in LSTM network are a) forget gate (f), b) input gate (i), control gate (c) and output gate (o) [19].

The underlying data pattern can be extracted and remembered, which addresses long-term data dependence on classic RNN algorithms [19]. Fig. 1 shows the TLSTM architecture [20]. Inputs of the architecture are h_{t-1} , x_t , and b . The term h_{t-1} represents previous cell state, x_t represents current input vector and b represents bias. One of the outputs of the architecture is c_t , which represents the present memory content. Another output of the architecture is h_t , represents present cell state. These four gates listed above influences the data in the memory cell. Forget gate gives a value in the range 0-1. This value defines how much should be ignored from the previous memory cell. If the forget gate produces a value close to 0 it means that at the new time stamp, much of the previous timestamp's memory will be overlooked and the reverse occurs for the value close to 1. The gates in TLSTM are represented in the following equations as follows:

Equation 1 represents the forget gate of TLSTM [20] as,

$$f_t = \alpha_g(w_f x_t + u_f h_{t-1} + b_f) \tag{1}$$

Equation 2 represents the input gate of TLSTM as,

$$i_t = \sigma_g(w_i x_t + u_i h_{t-1} + b_i) \tag{2}$$

Equation 3 represents the control gate of TLSTM as,

$$c_t = f_t \times c_{t-1} + i_t \times \sigma_h(w_c x_t + u_c h_{t-1} + b_c) \tag{3}$$

Equations 4 and 5 represent the output of TLSTM,

$$o_t = \sigma_g(w_o x_t + u_o h_{t-1} + b_o) \tag{4}$$

$$h_t = o_t \times \sigma_h c_t \tag{5}$$

Here, the sigmoid function is represented by σ_g and hyperbolic tangent function is denoted by σ_h . The symbols w and u represent weights. These weights usually prevent the issue of gradients from disappearing.

We have used 50 T-LSTM units in each layer. In each layer, for every input an attention value is calculated. Attention value gives the significance of the input and is helpful in final prediction. The dense layer allows the final prediction of whether a patient has diabetes with the aid of an attention vector.

It can be noticed from in Fig. 1, that there is no correlation between the previous memory content with any of the gates in the network. This results in an abnormal situation if the output gate is locked. This reduces the efficiency of prediction and classification tasks. Hence, the primary goal of this work is to apply CLSTM to the classification of patients with diabetes and to illustrate how CLSTM overcomes the limitations faced by TLSTM.

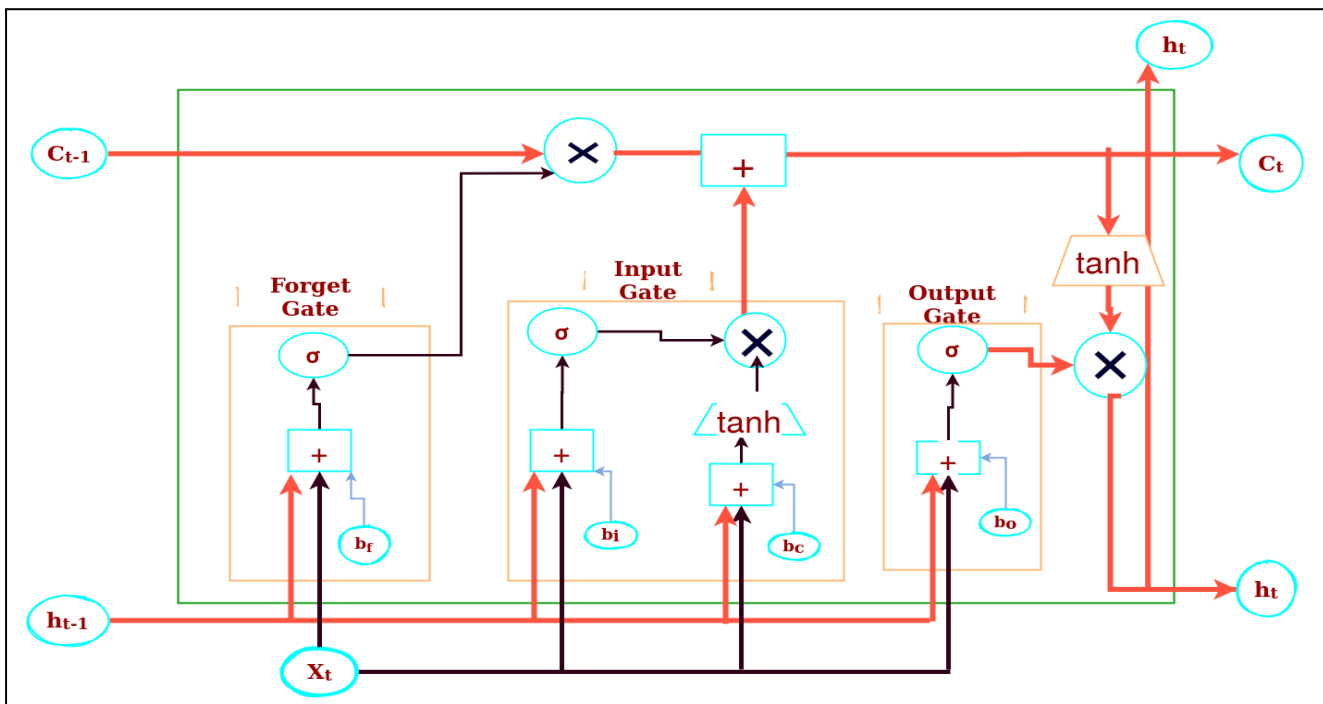


Fig. 1. Architecture of Traditional LSTM.

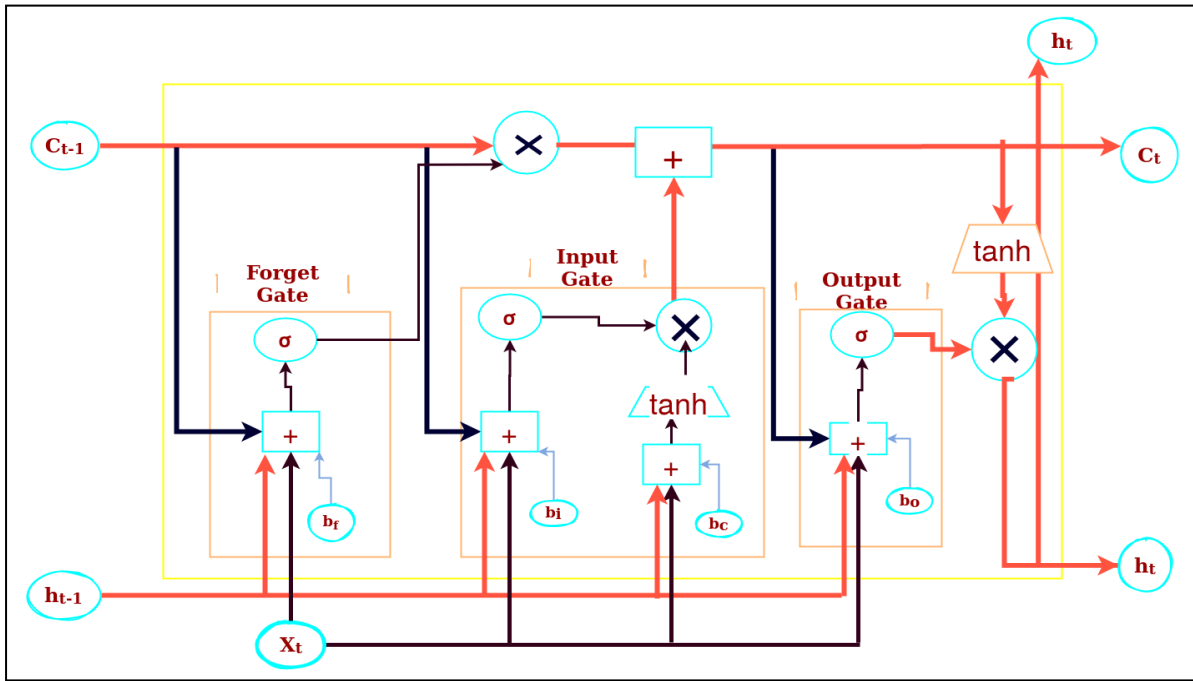


Fig. 2. Architecture of Convolutional LSTM.

B. Convolutional LSTM

Traditional LSTM do not access previous memory cell contents [19] even when the output gate of the model is closed. CLSTM negates this by adding an extra link to all the other gates from the previous memory. Fig. 2 [20] shows CLSTM diagram and its operation.

An additional parameter (former c_{t-1} memory content) is in CLSTM compared to TLSTM to provide previous memory cells' impact even when the output gate is closed.

The CLSTM four gates function with the help of the following equations:

Equation 6 represents the forget gate of CLSTM as,

$$f_t = \sigma_g(w_f x_t + u_f h_{t-1} + v_f c_{t-1} + b_f) \quad (6)$$

Equation 7 represents the input gate of CLSTM as,

$$i_t = \sigma_g(w_i x_t + u_i h_{t-1} + v_i c_{t-1} + b_i) \quad (7)$$

Equation 8 represents the control gate of CLSTM as,

$$c_t = f_t c_{t-1} + i_t \sigma_h(w_c x_t + u_c h_{t-1} + b_c) \quad (8)$$

Equations 9 and 10 represents the output of CLSTM as

$$o_t = \sigma_g(w_o x_t + u_o h_{t-1} + v_o c_{t-1} + b_o) \quad (9)$$

$$h_t = o_t \times \sigma_h(c_t) \quad (10)$$

This article developed a CLSTM-based model of diabetes prediction and is tested on the Pima Indian Diabetes dataset.

C. Proposed Model

Fig. 3 represents the proposed model for diabetes prediction. Initially, the PIMA dataset is pre-processed and

later, essential features are selected. For evaluation and training purposes, the dataset is split into train and test sets.

The hyperparameters of TLSTM and CLSTM models are tuned in the next step. Once the training phase on the dataset is completed, we have calculated various parameters for performance evaluation. The accuracy for different test sizes are measured in the paper in order to estimate the performance of the proposed model.

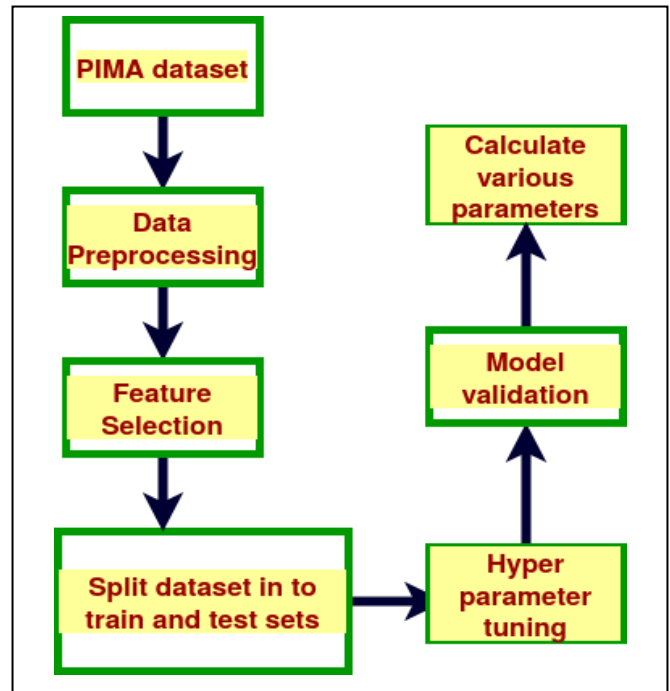


Fig. 3. Proposed Model for Diabetes Prediction.

D. Dataset Description

The initial process of our strategy is to apply dataset pre-processing techniques on the PIMA dataset. The dataset contains information about 768 patients, with nine attributes obtained for each patient. The data in the dataset consists of different female individuals between the ages of 21 and 81.

Six attributes represent physical examination specifics in each row, and the remaining attributes represent chemical examination information. The last attribute in-row is the data on whether the patient is diabetic.

The last column of each row is either 1 or 0, 1 indicating that the patient is diabetic and 0, indicating that the patient is not diabetic.

The first column in the dataset represents the number of times a woman is pregnant, and the second column in the dataset represents the plasma glucose concentration. The third column in the dataset depicts the diastolic blood pressure and the fourth column gives the thickness of the triceps skin fold. The fifth column represents serum insulin for two hours, and the sixth column represents the person's body mass index (BMI). Pedigree feature is in the seventh column and the eighth column in the dataset reflects the individual's age and the last

column reflects the incidence of diabetes (1/0). Fig. 4 shows the information of various attributes of the PIMA dataset. Fig. 5 shows the correlation of various attributes in the PIMA dataset.

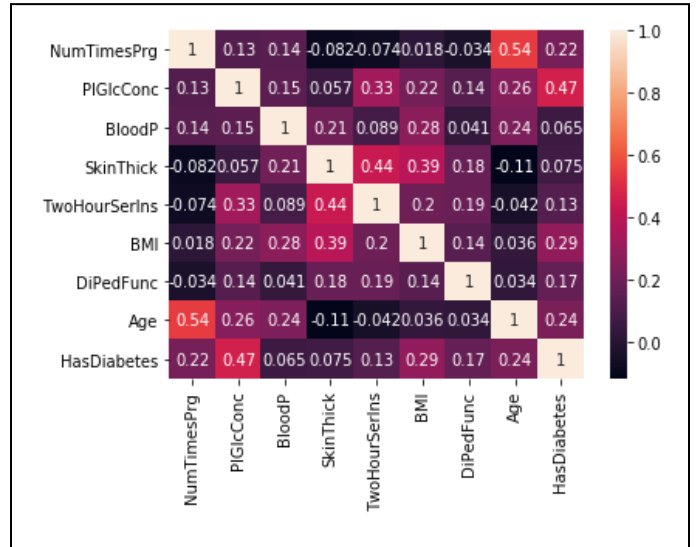


Fig. 4. Correlation of Various Attributes in the PIMA Dataset.

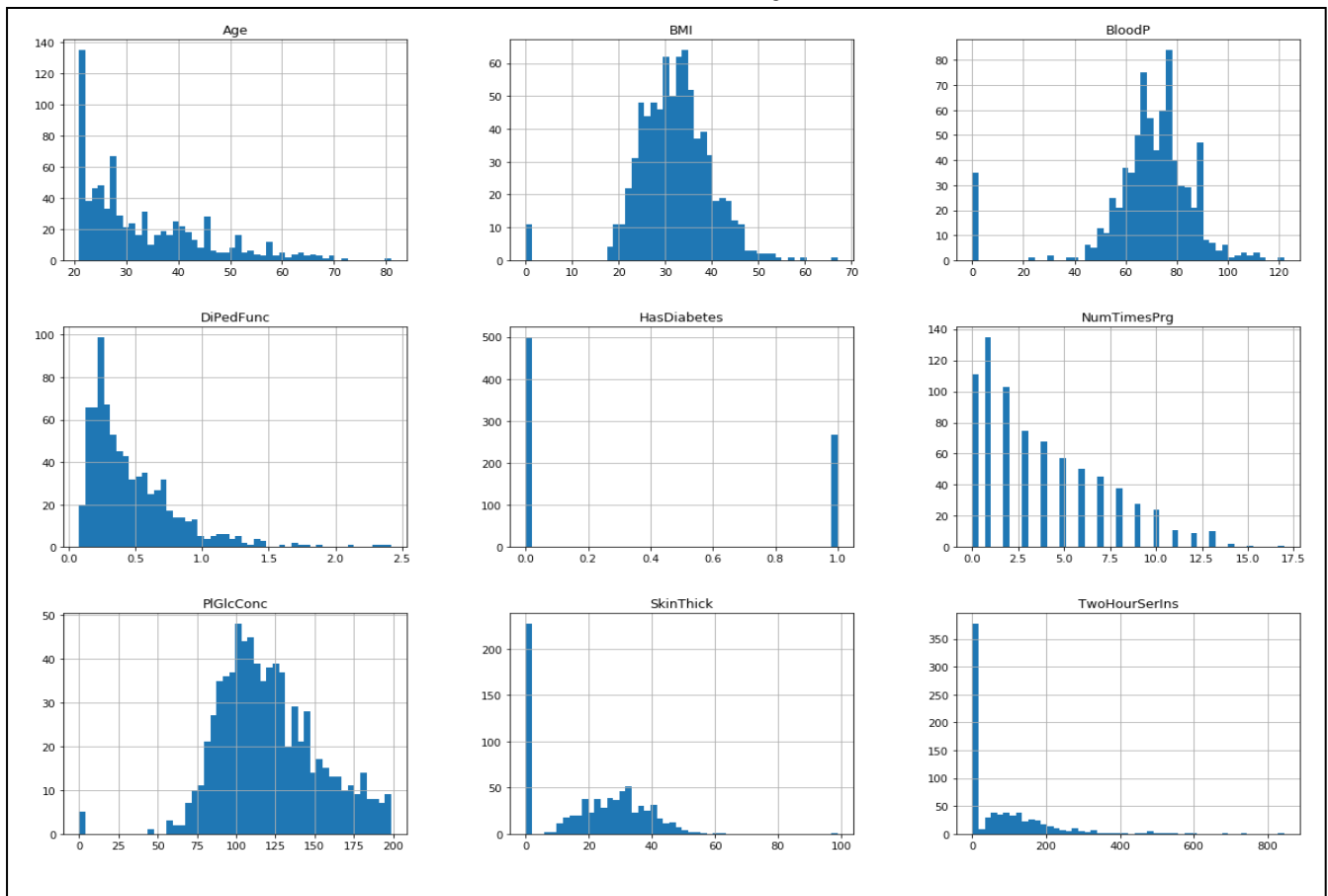


Fig. 5. Attributes in the PIMA Dataset.

IV. RESULTS AND DISCUSSION

The comparison of various models like neural networks, machine learning and deep learning systems are presented in this section.

A. Experimental Setup

The TLSTM, CLSTM models are used in this paper to predict feature selection. Initially we have pre-processed the dataset with the mentioned techniques in the previous section. Random Forest algorithm is used for feature selection. We have found from our observation that five features (Glucose, Age, BMI, BP, Insulin) as important.

We have set the TLSTM and CLSTM models' hyperparameters with the following details mentioned in Table 1.

The values in Table 1 are hyperparameter optimization values where we obtained highest accuracy. We have used python inbuilt packages to develop our model. Pre-processing and feature selection of dataset are also carried out using python.

Table 2 presents the results of different models on the PIMA dataset. Naïve Bayes, SVM, DT, K-means have similar accuracy results. TLSTM and CLSTM models outperformed the accuracy results of other existing models. The machine learning algorithms reported in this section are traditional ones.

In Table 2 all the results are obtained from our experimentations. The results show that the TLSTM and CLSTM models outperformed all the existing machine learning models.

TABLE I. HYPERPARAMETERS OF TLSTM AND CLSTM

Parameter	TLSTM	CLSTM
Learning Rate	0.02	0.01
Batch size	32	32
Hidden layers	50	50
Epoch	50	50

TABLE II. COMPARISON OF VARIOUS MODELS ON PIMA DATASET

Model	Accuracy (test set 10%)	Accuracy (test set 20%)
Naïve Bayes	79.6%	78.6%
SVM	79.2%	78%
Decision Trees	78.4%	77.2%
MLP	80%	82%
K means	77%	72%
TLSTM	92.5%	93.7%
CLSTM	96.8%	95.6%

The results presented in this section specify that the proposed model outperforms all the existing models. The TLSTM and CLSTM models have obtained higher accuracy results than all the existing machine learning models. The machine learning models do not capture the features properly and hence the results are less when compared with the proposed model. Moreover the proposed model takes care of the data pre-processing and feature selection properly and hence the results are high for our model.

V. CONCLUSION

This paper aims to introduce a CLSTM, TLSTM prediction model for diabetes. As diabetes is becoming a serious disorder now-a-days it is the need of the hour if the researchers come up with prediction models. The proposed approach enhances diabetes prediction using deep learning techniques. Moreover, the proposed approach also uses an efficient pre-processing mechanism called multivariate imputation by chained equations. This paper examines various classification approaches on the PIMA dataset. Existing ML and DL approaches are tested on PIMA dataset. As mentioned in Table 2, the result achieved by CLSTM model is higher than other methodologies. In the future, in the form of an application or a website, we plan to build a comprehensive framework using CLSTM algorithm, which will help practitioners to predict diabetes at early stages and reduce the risk of various diseases

REFERENCES

- [1] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. D. R. Fernandes, A. Ohlrogge and B. Malanda, "IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes Res. Clin. Pr.*, vol. 138, pp. 271–281, 2018.
- [2] Y. L. Sun and D. L. Zhang, "Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey," *Teh. Vjesn.*, vol. 26, pp. 872–880, 2019.
- [3] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders*, vol.19(1), pp.391-403, 2020.
- [4] D. Liccardo, A. Cannavo, G. Spagnuolo, N. Ferrara, A. Cittadini, C. Rengo and G. Rengo, "Periodontal disease: A risk factor for diabetes and cardiovascular disease," *International journal of molecular sciences*, vol. 20(6), pp.1414, 2019.
- [5] B. P. Nguyen, H. N. Pham, H. Tran, N. Nghiem, Q. H. Nguyen, T. T. Do, C.T. Tran and C. R. Simpson, "Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records," *Computer methods and programs in biomedicine*, vol. 182, 2019.
- [6] S. Spanig, A. Emberger-Klein, J. P. Sowa, A. Canbay, K. Menrad, and D. Heider, "The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes," *Artificial intelligence in medicine*, vol. 100, 2019.
- [7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and structural biotechnology journal*, vol. 15, pp.104-116, 2017.
- [8] J. P. Kandhasamy and S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus," *Procedia Comput. Sci.*, vol. 47, pp. 45–51, 2015.

- [9] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in medicine Unlocked*, vol. 17, 2019.
- [10] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," In *Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO)*, pp. 378–382, 2015.
- [11] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.
- [12] A. Negi and V. Jaiswal, "A first attempt to develop a diabetes prediction method based on different global datasets," In *Proceedings of the 2016 Fourth International Conference on Parallel, Distributed and Grid Computing*, pp. 237–241, 2016.
- [13] A. Ashiqzaman, A. Kawsar Tushar, M. D. Rashedul Islam, D. Shon, L. M. Kichang, P. Jeong-Ho, L. Dong-Sun and K. Jongmyon, "Reduction of overfitting in diabetes prediction using deep learning neural network," In *IT Convergence and Security; Lecture Notes in Electrical Engineering; Springer*, vol. 449, 2017.
- [14] G. Swapna, K. P. Soman and R. Vinayakumar, "Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals," *Procedia Comput. Sci.*, vol. 132, pp.1253–1262, 2018.
- [15] A. Mohebbi, T. B. Aradóttir, A. R. Johansen, H. Bengtsson, M. Fraccaro and M. Mørup, "A deep learning approach to adherence detection for type 2 diabetics," *IEEE Engineering in Medicine and Biology Society*, pp. 2896–2899, 2017.
- [16] R. Miotto, L. Li, B. A. Kidd and J.T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," *Appl. Sci.*, vol.6, pp. 4604–4612, 2019.
- [17] T. Pham, T. Tran, D. Phung and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Inform.*, vol.69, pp.218–229, 2017.
- [18] H. Balaji, N. Iyengar and R. D. Caytiles, "Optimal Predictive analytics of Pima Diabetics using Deep Learning," *Int. J. Database Theory Appl.*, vol. 10, pp. 47–62, 2017.
- [19] G. Zhu et al., "Redundancy and Attention in Convolutional LSTM for Gesture Recognition," *IEEE Trans. neural networks Learn. Syst.*, Jun. 2019.
- [20] G. Zhu, L. Zhang, L. Yang, L. Mei, S. A. A. Shah, M. Bennamoun, and P. Shen, "Redundancy and attention in convolutional LSTM for gesture recognition," *IEEE transactions on neural networks and learning systems*, vol. 31(4), pp.1323-1335, 2019.
- [21] Rahman and Siddiqui, "An Optimized Abstractive Text Summarization Model Using Peephole Convolutional LSTM," *Symmetry (Basel)*, vol. 11, 2019.

Identification of Babbitt Damage and Excessive Clearance in Journal Bearings through an Intelligent Recognition Approach

Joel Pino Gómez¹, Fidel E. Hernández Montero²
Technological University of Havana (Cujae)
Havana, Cuba

Julio C. Gómez Mancilla³, Yenny Villuendas Rey^{4*}
Instituto Politécnico Nacional (IPN)
CDMX, México

Abstract—Journal bearings play an important role on many rotating machines placed on industrial environments, especially in steam turbines of thermoelectric power plants. Babbitt damage (BD) and excessive clearance (C) are usual faults of steam turbine journal bearings. This paper is focused on achieving an effective identification of these faults through an intelligent recognition approach. The work was carried out through the processing of real data obtained from an industrial environment. In this work, a feature selection procedure was applied in order to choose the features more suitable to identify the faults. This feature selection procedure was performed through the computation of typical testors, which allows working with both quantitative and qualitative features. The classification tasks were carried out by using Nearest Neighbors, Voting Algorithm, Naïve Associative Classifier and Assisted Classification for Imbalance Data techniques. Several performance measures were computed and used in order to assess the classification effectiveness. The achieved results (e.g., six performance measures were above 0.998) showed the convenience of applying pattern recognition techniques to the automatic identification of BD and C.

Keywords—Journal bearing; Babbitt damage; excessive clearance; fault identification; feature selection; supervised classification

I. INTRODUCTION

Many rotating machine failures are related to bearing faults [1-5]. Journal bearings (JB) are usually found in heavy industries that include large rotating machines, whose early fault identification can yield a favorable impact on plant availability [6-9]. The stable operation of JB requires the clearance not to exceed the operating boundaries. If the clearance is out of bounds, mechanical instabilities in the shaft rotation, such as oil whirl, oil whip, looseness, or journal-to-bearing contact can arise [10, 11]. Mechanical stress due to these instabilities can also cause damage on the bearing babbitt surface, particularly, the oil whip is an unsafe operation that may cause severe damage on the machine [9-11]. Journal bearings are inspected during the maintenance process by a clearance measuring procedure, as well as an accurate examination of the babbitt surface; if a high-level damage occurs then a re-babbitting procedure will be necessary. Babbitt damage (BD) and excessive clearance (C) have been widely addressed in several researches [9-13].

Several classification methods have been applied on the automatic diagnosis of JB faults [14-33]. However, no research work addressing the automatic detection of BD and C in cylindrical journal bearings, through data gathered from real industrial environments, have been found by the authors of this paper, despite cylindrical journal bearings (CJB) are among the most common types of hydrodynamic journal bearings used by the turbomachinery [34, 35].

Even though the staff specialized in the diagnostic of JB faults evaluates a wide range of features expressed by numerical, ordinal and nominal variables (mixed features) [36-38], most research works addressing JB faults use only numerical variables, which are mainly vibration features extracted from both time and frequency domains [14-33]. Expert knowledge was previously considered in [38], but such a work only addressed the feature selection task.

This paper is focused on the automatic identification of BD and C in CJB by means of the processing of mixed features extracted from data gathered at a real industrial environment. The proposed methodology involves a feature selection procedure as a primary step, and then, the application of several classifiers. The mixed features processing is provided by the application of the Logical Combinatorial Pattern Recognition approach (LCPR) [39].

This paper is organized as follows: Section II provides a brief summary of previous works on the automatic fault diagnosis of JB. Section III presents some concepts and tools of the LCPR approach. The main features of the JB vibration spectrum that are traditionally used for diagnostics purposes are presented in Section IV. Section V presents the features and datasets used in this work. Section VI describes the proposed methodology. Section VII shows the main results and related discussions, and Section VIII presents the conclusions and future works.

II. REVIEW OF PREVIOUS WORKS ON AUTOMATIC FAULT DIAGNOSIS OF JOURNAL BEARINGS

Both supervised and unsupervised classification methods have been applied on the automatic diagnosis of JB faults [14-33]. While most of such works have used either data collected from a testbench [16-31] or data obtained from numerical models of faults [32, 33], just a few has used data gathered from a real life environment [14, 15]. For example, in [32, 33],

*Corresponding Author

different conditions of ovalization and wear were simulated by using numerical models. The classification methods used in these works were based on Convolutional Neural Networks (CNN) and resulted in a good accuracy. In [16-18], several faults (related to oil supply, looseness and bearing surface damages), induced in a journal bearing testbench, were diagnosed by means of Artificial Neural Networks (ANN) and Deep Neural Networks (DNN); a high effectiveness was achieved. Different friction and wear conditions were diagnosed in [24-26]. In these works, two test rigs were used and both Random Forest Classifier (RFC) and Support Vector Machines (SVM) were successfully applied. In [19-23], the effective diagnostic of unbalance, misalignment, rubbing and oil whirl was performed by applying Fisher Discriminant Analysis (FDA), Multilayer Perceptron (MLP), CNN and SVM. The data was gathered from a Bently-Nevada RK4 rotor kit and a feature selection procedure was implemented through the application of the Fisher Discriminant Ratio (FDR), Deep Belief Network (DBN), Kullback-Leibler Divergence (KLD) and Probability of Separation (PoS). In [27], a CNN classifier was applied on contact rubbing, block looseness, rotor unbalance and misalignment diagnostics. The results were compared with those obtained by SVM and a Probabilistic Neural Network (PNN). In that work, two testbenches were used. In [28], Genetic algorithms (GA) and ANN were used in order to identify three different lubrication conditions induced in a journal bearing test bed. The inadequate lubrication, oil starvation, corrosion, metal-to-metal contact, and extreme wear in the main journal bearing of an internal combustion engine were the faults addressed in [29-31]. In these works, k-Nearest Neighbor (kNN), Fisher Linear Discriminant (FLD), ANN and SVN classifiers were satisfactory applied. In [15], DBN, MLP, FDA and Self-Organizing Map (SOM) were the techniques used in order to diagnose the misalignment, the rubbing and the oil whirl produced in both a Bently-Nevada RK4 rotor kit and the journal bearings of a 500 MW steam turbine in a power plant. Several malfunctions like friction, abnormal lubrication and C were accurately diagnosed by means of Linear Discriminant Analysis (LDA), FDA and SVM techniques [14]. Such work was performed on induction motors and generators under full load conditions.

However, none of the previously mentioned paper address the automatic detection of BD and C in CJB, through data gathered from real industrial environments.

III. BRIEF INTRODUCTION TO LCPR APPROACH

LCPR constitutes an approach suitable to deal with mixed data (i.e., both quantitative and qualitative features) in feature selection and pattern classification applications. This approach provides several useful tools for processing mixed and incomplete data [39]. LCPR involves multiples comparison criteria to establish comparisons between the values of a feature. A comparison criterion (CC) is a mathematical formulation that allows for computing the similarity or dissimilarity between the values taken by a feature for two different objects. The following CCs are two examples that allow for determining the dissimilarity between either nominal or numerical features, respectively:

$$CC_1(X_s(O_i), X_s(O_j)) = \begin{cases} 1 & \text{if } X_s(O_i) \neq X_s(O_j) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$CC_2(X_s(O_i), X_s(O_j)) = \begin{cases} 1 & \text{if } |X_s(O_i) - X_s(O_j)| > \sigma_s \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $X_s(O_i)$ and $X_s(O_j)$ are the values of the feature s for the objects O_i and O_j , respectively, and σ_s is the standard deviation of the values taken by the feature s , in case of being a numerical feature. The output takes values '0' or '1' indicating that the comparison results are similar or dissimilar, respectively. For example, Table 1 presents three objects described by two features: the feature 1 (a nominal feature) and the feature 2 (a numerical feature).

Then, CC1 can be used in order to compare the values taken by feature 1 and CC2 can be used for the values taken by feature 2. Assuming that the standard deviation of feature 2 is $\sigma_2=0.28$, the comparison (Cr) between the three objects results yields:

$$Cr = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where the row 1 is the result of comparing O_1 and O_2 : "dissimilar" for both features. The row 2 is the result of comparing O_1 and O_3 : "similar" for feature 1 and "dissimilar" for feature 2. And the row 3 is the result of comparing O_2 and O_3 : "dissimilar" for feature 1 and "similar" for feature 2.

The feature selection can be accomplished by means of a useful tool: The Typical Testor (TT) computation. A testor (T) is defined as a subset of features that allows for differentiating between any two objects that belong to different classes; CCs defined for the comparison of such features are used. A TT is an irreducible T; that is, if any feature of a TT is removed, then the TT stops being a T [39]. Therefore, a TT is the most compact form in which a testor can appear. In a pattern recognition problem, the set of all TTs contains all the minimum-length subsets of features that allow for class differentiation. The TTs make contribution to both the classification process and the selection of only the more significant features.

Sometimes, the computation of the whole set of TTs can take long times. This is due to the algorithms need to check several features subsets, bounded by the exponential of the number of features. Undoubtedly, this is a non-polynomial problem that could incur in a high computational cost. Several algorithms or methods have been developed for the minimization the TT searching time [40]. In this work, the TTs were computed by means of one of the most powerful algorithms: the fast-BR algorithm [41].

TABLE I. EXAMPLE OF OBJECT DESCRIPTION

Object	Feature 1	Feature 2
O_1	1Xh	3.07
O_2	2Xv	1.78
O_3	1Xh	2.03

According to the pattern classification task, testors bring out an idea about which features are more significant or which features provide more information. Accordingly, the importance of a feature can be assessed by the number of TTs that include such a feature [39]. That is,

$$P(x) = \frac{\omega(x)}{\omega} \quad (3)$$

where ω is the number of TTs and $\omega(x)$ is the number of TTs that include the feature x . According to equation (3), the higher the number of TTs that include a feature, the higher the importance of such a feature. In addition, the feature importance can be assessed through the dimensions of the TTs that include such a feature [39]. That is,

$$L(x) = \frac{\sum_{t \in \psi(x)} \frac{1}{t}}{\psi(x)} \quad (4)$$

where t is the number of features of each TT including the feature x , and $\psi(x)$ is the number of TTs that include the feature x . According to equation (4), a feature is more important as it is found in shorter TTs. Finally, the feature importance can be expressed as follows [39]:

$$\rho(x) = \alpha P(x) + \beta L(x) \quad (5)$$

where α and β are weighting coefficients of $P(x)$ and $L(x)$, respectively. Then, the features selection is completed by removing the resulting features with importance values below thresholds empirically established.

IV. FEATURES OF THE JOURNAL BEARING VIBRATION SPECTRUM

Vibration analysis is essential for the condition evaluation of journal bearings [37]. Frequency domain representations of vibration signals bring out features very significant to JB fault diagnosis [21]. The spectrum of the journal bearing vibration signals usually exhibits several harmonics. If X is the value of the rotational speed in Hz, then some features of the vibration spectrum that are usually inspected for diagnostics purposes are [21, 37]: the synchronous spectral component (the amplitude of the spectral component at frequency $1X$), its harmonics (the amplitudes of the spectral components at frequencies corresponding to integer multiples of $1X$, e.g., $2X$, $3X$, ...), its inter-harmonics (the amplitudes of the spectral components between successive harmonics, e.g., $1.5X$), and its sub-synchronous (the amplitudes of the spectral components under frequency $1X$, e.g., $0.4X$).

Figure 1 shows an example of a real JB velocity vibration spectrum, where some of the aforementioned features can be seen. Taking into account that the rotational speed is 3600 revolutions per minute (60 Hz), the more predominant spectral components are the harmonics $1X$, $2X$, $3X$, $4X$, $6X$ and $8X$. These features could be used for automatic JB fault diagnosis [20, 21, 33]; however, in several cases many of them could either appear at very low amplitudes or be not visible in the spectrum (e.g., the sub-synchronous, inter-harmonics or some harmonics components do not appear in the spectrum shown in

Figure 1). In such cases, these features will not be contributing with meaningful information to the diagnosis process.

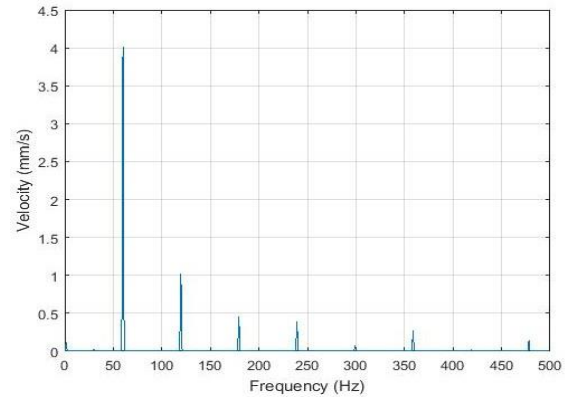


Fig. 1. Velocity Vibration Spectrum of a Real JB.

Other features, such as ratios between the aforementioned features, as well as different statistical measures, can also be extracted from the vibration spectrum [20, 31].

V. FEATURES AND DATASETS SUPPORTING THE CLASSIFICATION TASKS

From the set of features of the velocity vibration spectrum, presented in Section 3, this work addressed the use of the features that provided the information more useful to the diagnosis procedure. In this work, a new approach of feature arrangement was applied: while a small number of features is well defined, another group of features is undetermined and will be defined by the spectral components reaching the largest amplitudes. In this work, 35 mixed features were used for the pattern classification task; a brief description of them is presented in Table 2.

The feature 1 was the synchronous spectral component of the horizontal vibration; features from 2 to 6 were the highest-amplitude spectral components of the horizontal vibration, regardless the frequencies at which they were given rise; feature 7 was the synchronous spectral component of the vertical vibration; features from 8 to 12 were the highest-amplitude spectral components of the vertical vibration, regardless the frequencies at which they were given rise; features from 13 to 17 were the ratios of the highest-amplitude spectral components of the horizontal vibration (features from 2 to 6) to the synchronous spectral component of the horizontal vibration (feature 1); features from 18 to 22 were the ratios of the highest-amplitude spectral components of the vertical vibration (features from 8 to 12) to the synchronous spectral component of the vertical vibration (feature 7); feature 23 was the ratio of the synchronous spectral component of the horizontal vibration (feature 1) to the synchronous spectral component of the vertical vibration (feature 7); and features from 24 to 35 were the names (nominal features) of the spectral components denoted by features from 1 to 12 arranged in descendent order according to their values. The nominal features allow for the identification of the spectral components selected as the first 12 features and supply information about their amplitude order. Two examples of the set of features can be found in Table 3.

TABLE II. FEATURE DESCRIPTION

No.	Value description (H: Horizontal Vibration; V: Vertical Vibration)	Frequency	Domain (R: Real; N: Nominal)
1	amplitude of the synchronous component (H), 1Xh	Rotational	R
2	highest amplitude of a spectral component (H) different to 1Xh	?	R
3	second highest amplitude of a spectral component (H) different to 1Xh	?	R
⋮	⋮	⋮	⋮
6	fifth highest amplitude of a spectral component (H) different to 1Xh	?	R
7	amplitude of the synchronous component (V), 1Xv	Rotational	R
8	highest amplitude of a spectral component (V) different to 1Xv	?	R
9	second highest amplitude of a spectral component (V) different to 1Xv	?	R
⋮	⋮	⋮	⋮
12	fifth highest amplitude of a spectral component (V) different to 1Xv	?	R
13	rate of value of feature 2 to the value of feature 1	--	R
14	rate of value of feature 3 to the value of feature 1	--	R
⋮	⋮	⋮	⋮
17	rate of value of feature 6 to the value of feature 1	--	R
18	rate of value of feature 8 to the value of feature 7	--	R
19	rate of value of feature 9 to the value of feature 7	--	R
⋮	⋮	⋮	⋮
22	rate of value of feature 12 to the value of feature 7	--	R
23	rate of value of feature 1 to the value of feature 7	--	R
24	name of the feature from 1 to 12 with the highest amplitude	?	N
25	name of the feature from 1 to 12 with the second highest amplitude	?	N
⋮	⋮	⋮	⋮
35	name of the feature from 1 to 12 with the twelfth highest amplitude	?	N

TABLE III. TWO EXAMPLES OF SET OF FEATURES

Objects	Features Values
object 1	5.03, 1.61, 1.04, 1.04, 0.92, 0.76, 1.43, 1.84, 1.45, 1.24, 1.15, 0.78, 0.32, 0.21, 0.21, 0.18, 0.15, 1.29, 1.02, 0.87, 0.81, 0.55, 3.52, 1Xh, 3Xv, 2Xh, 2Xv, 1Xv, 4Xv, 5Xv, 6Xh, 3Xh, 4Xh, 6Xv
object 2	3.13, 0.51, 0.39, 0.32, 0.16, 0.12, 3.27, 1.71, 1.52, 1.27, 0.53, 0.53, 0.16, 0.13, 0.1, 0.05, 0.04, 0.52, 0.46, 0.39, 0.16, 0.16, 0.96, 1Xv, 1Xh, 2Xv, 4Xv, 3Xv, 6Xv, 5Xv, 6Xh, 2Xh, 3Xh, 5Xh, 4Xh

The dataset supporting this research were taken from diagnostics and maintenance reports of a 100 MW steam turbine running for three years in an active thermoelectric

power plant. We cannot disclose the name of the thermoelectric plant due to confidentiality issues. Data was collected from four journal bearings that were affected by five different fault conditions: C, BD in the bottom half of the journal bearing (B), BD in the top half of the journal bearing (T), faults B and C occurring simultaneously (BC), and faults B, C and T occurring simultaneously (BCT). Data corresponding to healthy condition of operation was not available. The total number of measurements (objects) to work with was 3314.

Each object description is expressed by the set of 35 features proposed and described in previous section. These features were extracted from the JB absolute vibration measurements taken at both horizontal and vertical directions, at a sampling frequency equal to 1 kHz. Such measurements were performed and stored by means of the online monitoring system installed at the thermoelectric power plant. The online monitoring system and the velocity sensor installed for the vibration acquisition were the VIBROCONTROL 4000 and BK VIBRO VS-079, respectively. Each measurement consisted of an 800-lines spectrum.

The object distribution is shown in Table 4. This table reveals that the object quantities were imbalanced for the five classes of faults, with an Imbalance Ratio IR=6.61.

TABLE IV. OBJECT DISTRIBUTION

Fault Conditions	Objects Numbers
B	431
C	708
T	213
BC	1409
BCT	553

VI. METHODOLOGY

The aim of this research is to identify automatically the BD and C in CJB through data taken from real industrial environments. Figure 2 shows a scheme of the methodology developed in order to accomplish this goal; this methodology is described as follows.

Firstly, a feature selection process addressing the determination of the TTs from the set of 35 features presented in the section 4, was implemented. The algorithm used in order to search the TTs was the fast-BR [41]. The feature importance was computed by applying the equation (5) for both parameters α and β being equal to 0.5. After the implementation of several tests, the threshold for feature selection was chosen to be the difference between the mean and the half of the standard deviation of the computed importance records.

Afterwards, a classification strategy was implemented: on one hand, the classification procedure was carried out by using only the important (selected) features and, on the other hand, the classification procedure was performed by using the whole set features. This strategy will reveal how effective the implemented feature selection procedure was.

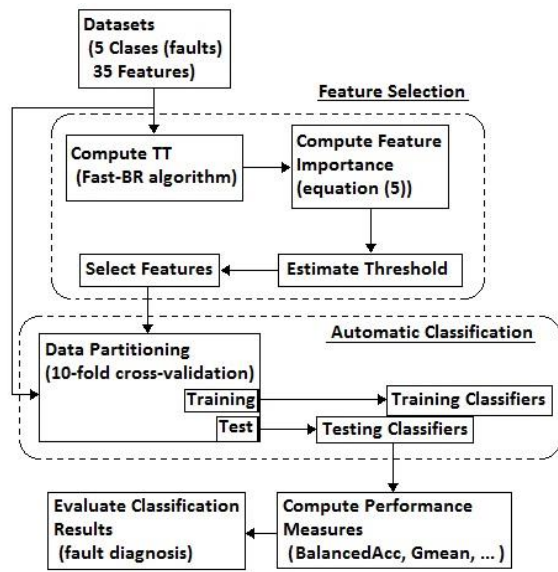


Fig. 2. Methodology.

Four classification techniques were applied: kNN [42], Voting Algorithm (ALVOT) [39], Naïve Associative Classifier (NAC) [43] and Assisted Classification for Imbalance Data (ACID) [44]. These methods are very suitable to be applied when mixed-data processing is required and they have exhibited high effectiveness in different scenarios. The initialization parameters and dissimilarity function of the ALVOT and NAC algorithms were the same as those presented in [43]. For the kNN and ACID classifiers the dissimilarity function used was HEOM [45]. Table 5 summarizes the parameters used for the compared classifiers.

The classification tests were carried out on the Experimental Platform for Intelligent Computing (EPIC) [46]. Although this platform is not among the most popular tools for intelligent computing experiments (such as WEKA [47] and KEEL [48]), it allows for processing mixed and incomplete data and it provides the classifiers proposed to be applied on this research.

Cross-validation methodology, specifically, the k-fold cross-validation procedure with k equal to 10, was applied in order to warrant the reliability of the results. This procedure has been widely employed in the context of pattern recognition, machine learning and data mining, and the most common scheme has been the 10-fold cross-validation [49]. Although the available dataset is imbalanced, the class with the lowest number of objects admits 10-fold cross-validation.

The performance measures applied for the evaluation of the classification results, given the imbalance of the dataset presented in Table 2, were: the balanced accuracy (BalancedAcc) [50], the geometric mean of the recall measure (Gmean) [51], the macro precision (PrecisionM) [52], the macro geometric mean of the precision and recall measures (GmeasureM) [53], the macro F-measure (FScoreM) [52] and the kappa (Kappa) statistic [54]. These indexes are good measures of the classifiers' performance and they are recommended to be used in multiclass and imbalanced problems [54].

TABLE V. CLASSIFIER PARAMETERS

Classifier	Parameter values
ACID	$N_p=25, it = 100, \varepsilon = 0.1$, Dissimilarity: HEOM
ALVOT	SSS: Typical testers, Decision rule: class with maximum $I_j^i(o)$, Similarity: 1/HEOM $I_j^i(o) = \frac{\sum_{n_i \in S} r_{n_i^i(o,y)}^j}{ S }$, $\Gamma_{n_i^j}^j(o) = \frac{\sum_{y \in T_j} r_{n_i^j(o,y)}^j}{ K_j }$, $\Gamma_{n_i}(o, y) = \rho_y * \rho_y * \beta(o, y, \Omega_i)$
kNN	$k = 1$, Dissimilarity: HEOM
NAC	$w_j = 1$ for all features

The comparison criteria given by equations (1) and (2) were used in order to compare both nominal and numerical features, respectively, during the feature selection and classification procedures.

VII. RESULTS AND DISCUSSION

The number of TTs obtained from the set of 35 features presented in the section 4 was 31218. The feature importance results are shown in Figure 3. In this figure, selected and removed features are shown through colors green and red, respectively. The decision threshold used in order to select the features is the red horizontal line shown in Figure 3. The number of selected and removed features was 23 (65.71%) and 12 (34.29%), respectively. There were no features with importance values equal to zero (each feature was found at least in one TT). These 23 selected features formed a set of features used in order to perform one of the two implemented fault classification procedures. Another fault classification procedure was carried out by using the set of features given by all of features.

The results obtained by the application of the four selected classifiers on the identification of the faults are shown in Figures 4 to 7 (gray color for the results obtained when the whole set of features was used; green color for the results obtained when the selected features were used).

The results obtained from the application of ALVOT classifier are shown in Figure 4. In general, the best performance was obtained when the set of selected features was used, except for the PrecisionM measure, which reached a value slightly higher for the set of all features. With regard to the six performance measures, the application of the ALVOT classifier yielded values higher than 82 % and according to the BalancedAcc and Gmean measures the values were higher than 90 %.

The results obtained by means of the NAC classifier are shown in Figure 5. According to the six performance measures, the best results were obtained when the set of selected features was used. The six performance measures show values higher than 92%, which proves that NAC is a good classifier for automatic identification of BD and C in CJB.

Figure 6 shows the results obtained by means of the kNN classifier. In this case, the best performance was obtained when the set of all features was used. kNN was an effective classifier for the automatic fault identification, since the six performance measures yielded values higher than 95%.

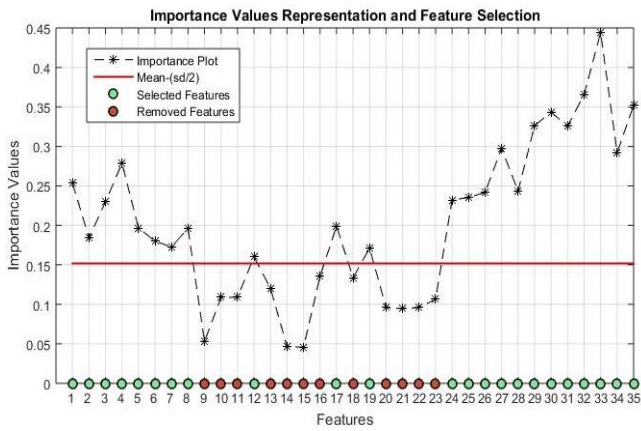


Fig. 3. Feature Importance for the Identification of B, C, T, BC and BCT Faults.

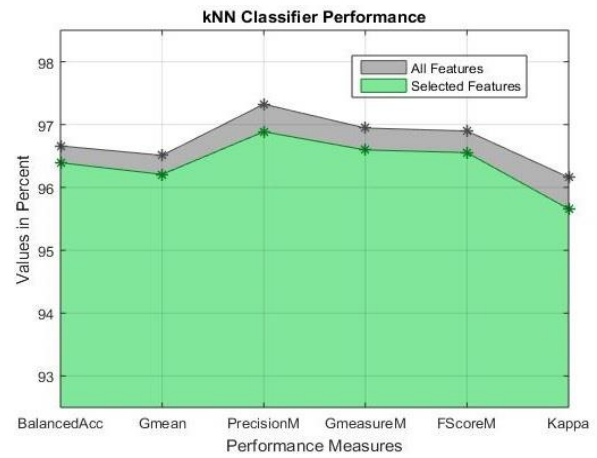


Fig. 6. Results obtained by kNN Classifier.

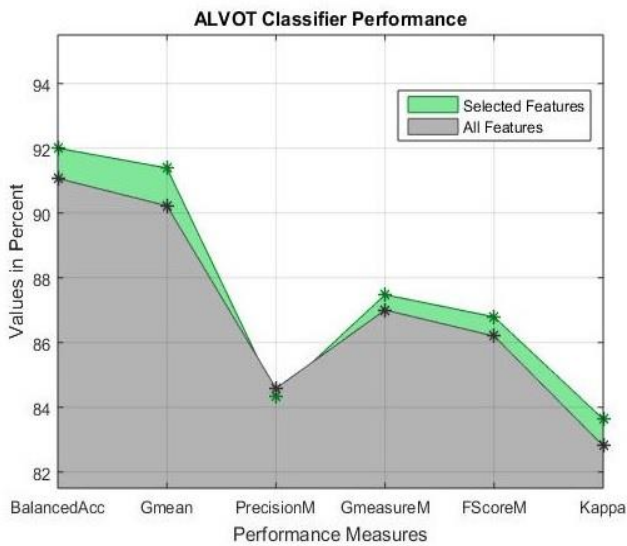


Fig. 4. Results obtained by ALVOT Classifier.

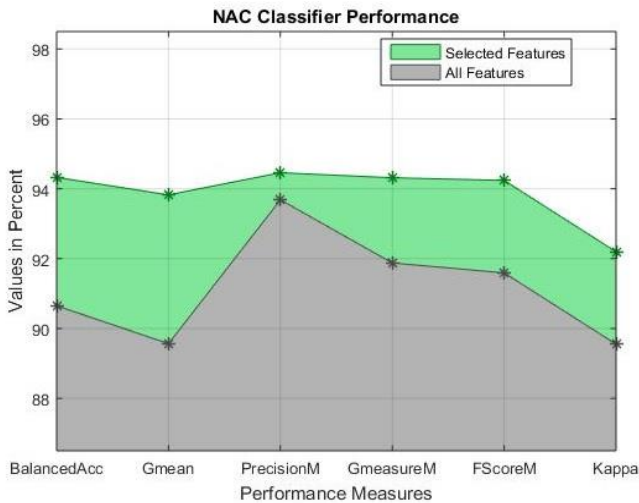


Fig. 5. Results obtained by NAC Classifier.

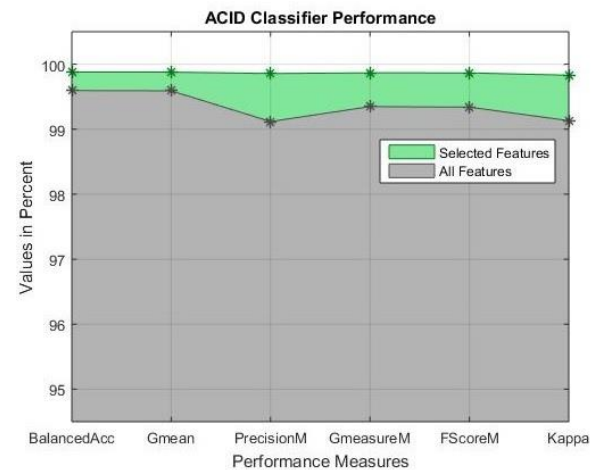


Fig. 7. Results obtained by ACID Classifier.

Figure 7 shows the results obtained when the ACID classifier was applied. In this case, as the six performance measures reveal, the best results were obtained when the set of selected features was used; every computed measure yielded values higher than 99.8%. This classifier also proved to be suitable for the automatic fault identification in CJB.

Summarizing, the values achieved by the computed performance measures proved that the four classifiers are suitable to be applied on fault diagnostics of CJB when the features proposed in this paper are used. It should be noticed that the kNN and ACID classifiers yielded the best performances.

These results validate the methodology proposed for the automatic identification of BD and C in CJB. This methodology involves the use of TTs for feature selection and the classifiers ALVOT, NAC, kNN and ACID for fault diagnosis.

These results are highly relevant since they constitute a first report on the automatic identification of BD and C in CJB through data collected at a real industrial environment. Besides, the importance of using mixed features for such purpose was also validated for first time.

VIII. CONCLUSIONS

This paper presents the results of a study about the identification of BD and C in journal bearing through the use of features extracted from the vibration spectrum. In this work, the faults and data processed have come from four journal bearings of a 100 MW steam turbine placed in an active thermoelectric power plant.

To the best of our knowledge, this work constitutes the first study addressing the automatic classification of BD and C in CJB placed in a real industrial environment. Besides, the use of only the more useful features that could be extracted from the vibration spectrum, as well as the use of both numerical and nominal features (all representing the expert's knowledge) for JB fault identification is proposed as a new methodological approach that led to remarkable results.

The classification process, which consisted in using four different classifiers and working with both selected features and the whole set of features, was very successful, since the effectiveness obtained was very high. In particular, the highest performance (99.8%) was achieved by ACID algorithm. This algorithm and kNN are the classifiers recommended to be used for the identification of BD and C in journals bearings. The search of typical testors is recommended for performing feature selection.

Several significant novelties were presented in this paper: the use of real-world dataset for CJB fault identification; the use of a new set of features, involving both numerical and nominal features for fault identification; the implementation of a feature selection procedure for improving the classification tasks; and the application of two effective classifiers (NAC and ACID) on the automatic fault diagnosis of machinery.

As future work we want to address other feature selection techniques, as well as other strategies for computing feature importance.

ACKNOWLEDGMENT

The authors would like to thank the support given by the EPIC research team, from IPN, Mexico, as well as the contributions of the diagnostic and maintenance experts: Julio González Martínez, Yuritz Cruz Guzmán, Jorge C. Arce Miranda and María Antonia Téllez. The authors thank the management of the thermoelectric power plant for providing access to the data used in accordance with the confidentiality agreement signed. The authors also thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, ESIME and CIDETEC), the CONACyT, and SNI for their economical support to develop this work.

REFERENCES

- [1] L. Ruonan, Y. Boyuan, Z. Enrico, and C. Xuefeng, "Artificial intelligence for fault diagnosis of rotating machinery: A review," *Mechanical Systems and Signal Processing*, vol. 108, pp. 33-47, 2018.
- [2] G. Królczyk, Z. Li, and J. A. Antonino Daviu, "Fault Diagnosis of Rotating Machine," *Applied Sciences*, vol. 10, no. 6, p. 4, 2020.
- [3] T. Qiang, P. Flores, and H. M. Lankarani, "A comprehensive survey of the analytical, numerical and experimental methodologies for dynamics systems with clearance or imperfect joints," *Mechanism and Machine Theory* vol. 122, pp. 1-57, 2018.
- [4] S. Schmidt, P. S. Heyns, and K. C. Gryllias, "A pre-processing methodology to enhance novel information for rotating machine diagnostics," *Mechanical Systems and Signal Processing*, vol. 124, pp. 541-561, 2019.
- [5] Y. Wei, Y. Li, M. Xu, and W. Huang, "A review of early fault diagnosis approaches and their applications in rotating machinery," *Entropy*, vol. 21, no. 4, p. 26, 2019.
- [6] J. Pino Gómez et al., "Maintenance importance of mechanical elements and faults in steam turbines. Data history analysis," *Ingeniería Energética*, vol. 38, no. 2, pp. 106-114, 2017.
- [7] N. Ding, H. Li, Z. Yin, and F. Jiang, "A novel method for journal bearing degradation evaluation and remaining useful life prediction under different working conditions," *Measurement*, vol. 177, 2021.
- [8] H. Al-Mosawy, H. Jamali, and M. Tolephih, "Effects of linear modification on the performance of finite length journal bearings," in *IOP Conference Series: Materials Science and Engineering*, 2021.
- [9] L. Zhang, H. Xu, S. Zhang, and S. Pei, "A radial clearance adjustable bearing reduces the vibration response of the rotor system during acceleration," *Tribology International*, vol. 144, p. 15, 2020.
- [10] J. Junyeong et al., "Monitoring Journal-Bearing Faults: Making Use of Motor Current Signature Analysis for Induction Motors," *IEEE Industry Applications Magazine*, vol. 23, no. 4, pp. 12-21, 2017.
- [11] L. B. Visnadi and H. F. de Castro, "Influence of bearing clearance and oil temperature uncertainties on the stability threshold of cylindrical journal bearings," *Mechanism and Machine Theory*, vol. 134, pp. 57-73, 2019.
- [12] K. K. Yadav et al., "Studies and Analysis of Effect of Foreign Particles on the Parts of Steam Turbine," *International Journal of Applied Engineering Research*, vol. 13, no. 6, pp. 386-395, 2018.
- [13] R. Ranjan, S. K. Ghosh, and M. Kumar, "Fault diagnosis of journal bearing in a hydropower plant using wear debris, vibration and temperature analysis: A case study," *Journal of Process Mechanical Engineering*, vol. 234, no. 3, pp. 235-242, 2020.
- [14] Y. Elyassami, K. Benjelloun, and M. E. Aroussi, "Sleeve Bearing Fault Diagnosis and Classification," *WSEAS Transactions On Signal Processing*, vol. 12, pp. 2224-3488, 2016.
- [15] H. Oh, J. H. Jung, B. C. Jeon, and B. D. Youn, "Scalable and Unsupervised Feature Engineering Using Vibration-Imaging and Deep Learning for Rotor System Diagnosis," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 4, pp. 3539-3549, 2017.
- [16] T. Narendiranath Babu, H. S. Himamshu, K. N. Prabin, P. D. Rama, and C. Nishant, "Journal Bearing Fault Detection Based on Daubechies Wavelet," *Archives of Acoustics*, vol. 42, no. 3, pp. 401- 414, 2017.
- [17] T. Narendiranath Babu, T. Manvel Raj, and T. Lakshmanan, "Application of Butterworth filter for fault diagnosis on journal bearing," *Journal of VibroEngineering*, vol. 16, no. 3, pp. 1602-1617, 2014.
- [18] T. Narendiranath Babu, A. Aravind, A. Rakesh, M. Jahzan, D. Rama Prabha, and M. R. Viswanathan, "Automatic Fault Classification for Journal Bearings Using ANN and DNN," *Archives of Acoustics*, vol. 43, no. 4, pp. 727-738, 2018.
- [19] B. C. Jeon, "Statistical Approach to Diagnostic Rules for Various Malfunctions of Journal Bearing System Using Fisher Discriminant Analysis," in *European Conference of the prognostics and health management society*, 2014.
- [20] J. H. J. Byung Chul Jeon, Byeng Dong Youn, Yeon-Whan Kim, and Yong-Chae Bae, "Datum Unit Optimization for Robustness of a Journal Bearing Diagnosis System," *International Journal of Precision Engineering and Manufacturing*, vol. 16, no. 11, pp. 2411-2425, 2015.
- [21] J. Joon Ha, J. Byung Chul, Y. D. Byeng , K. Myungyon, K. Donghwan, and K. Yeonwhan, "Omnidirectional regeneration (ODR) of proximity sensor signals for robust diagnosis of journal bearing systems," *Mechanical Systems and Signal Processing*, vol. 90, pp. 189-207, 2017.
- [22] H. Oh, B. C. Jeon, J. H. Jung, and B. D. Youn, "Smart diagnosis of journal bearing rotor systems: Unsupervised feature extraction scheme by deep learning," in *Annual Conference of the Prognostics and Health Management Society*, 2016.
- [23] B. C. Jeon, J. H. Jung, M. Kim, K. H. Sun, and B. D. Youn, "Optimal vibration image size determination for convolutional neural network

- based fluid-film rotor-bearing system diagnosis," *Journal of Mechanical Science and Technology*, vol. 34, no. 4, pp. 1467-1474, 2020.
- [24] N. Mokhtari and C. Gühmann, "Classification of journal bearing friction states based on acoustic emission signals," *tm-Technisches Messen*, vol. 85, no. 6, pp. 434-442, 2018.
- [25] N. Mokhtari, J. G. Pelham, S. Nowoisky, J.-L. Bote-Garcia, and C. Gühmann, "Friction and Wear Monitoring Methods for Journal Bearings of Geared Turbofans Based on Acoustic Emission Signals and Machine Learning," *Lubricants*, vol. 8, no. 3, p. 27, 2020.
- [26] J.-L. Bote-Garcia, N. Mokhtari, and C. Gühmann, "Wear monitoring of journal bearings with acoustic emission under different operating conditions," in *PHM Society European Conference*, 2020.
- [27] S. Guo, T. Yang, W. Gao, and C. Zhang, "A Novel Fault Diagnosis Method for Rotating Machinery Based on a Convolutional Neural Network," *Sensors*, vol. 18, no. 5, pp. 1429-1444, 2018.
- [28] S. Hosseini, M. Ahmadi Najafabadi, and M. Akhlaghi, "Classification of acoustic emission signals generated from journal bearing at different lubrication conditions based on wavelet analysis in combination with artificial neural network and genetic algorithm," *Tribology International*, vol. 95, pp. 426-434, 2016.
- [29] A. Moosavian, H. Ahmadi, and A. Tabatabaefar "Fault Diagnosis of main engine journal bearing based on vibration analysis using Fisher linear discriminant, K-nearest neighbor and support vector machine," *Journal of Vibroengineering* vol. 14, no. 2, pp. 894-906, 2012.
- [30] A. Moosavian, H. Ahmadi, A. Tabatabaefar, and B. Sakhaei, "An Appropriate Procedure for Detection of Journal-Bearing Fault Using Power Spectral Density, K-Nearest Neighbor and Support Vector Machine," *International Journal on Smart Sensing and Intelligent Systems*, vol. 5, no. 3, pp. 685-700, 2012.
- [31] A. Moosavian, "Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing," *Shock and Vibration*, vol. 20, no. 2, pp. 263-272, 2013.
- [32] D. Stuardi Alves et al., "Uncertainty quantification in deep convolutional neural network diagnostics of journal bearings with ovalization fault," *Mechanism and Machine Theory*, vol. 149, 2020.
- [33] O. Gecgel et al., "Simulation-Driven Deep Learning Approach for Wear Diagnostics in Hydrodynamic Journal Bearings," *Journal of Tribology*, vol. 143, no. 8, p. 9, 2020.
- [34] S. M. DeCamilo, A. Dadouche, and M. Fillon, "Journal Bearings in Power Generation," in *Encyclopedia of Tribology*: Springer, 2013.
- [35] S. Chatterton, P. Vinh Dang, P. Pennacchi, A. De Luca, and F. Flumian, "Experimental evidence of a two-axial groove hydrodynamic journal bearing under severe operation conditions," *Tribology International*, vol. 109, pp. 416-427, 2017.
- [36] A. Muszynska, "Vibrational Diagnostics of Rotating Machinery Malfunctions," *International Journal of Rotating Machinery*, vol. 1, no. 3-4, pp. 237-266, 1995.
- [37] A. Bilošová and J. Biloš, *Vibrations Diagnostics*. Ostrava: VSB - Technical University of Ostrava, 2012.
- [38] J. Pino Gómez, F. E. Hernández Montero, and J. C. Gómez Mancilla, "Variable Selection for Journal Bearing Faults Diagnostic Through Logical Combinatorial Pattern Recognition," in *Lecture Notes in Computer Science*: Springer, 2018.
- [39] J. Ruiz-Shulcloper, "Pattern Recognition with Mixed and Incomplete Data," *Pattern Recognition and Image Analysis*, vol. 18, no. 4, pp. 563-576, 2008.
- [40] V. Rodríguez-Diez, J. F. Martínez-Trinidad, M. S. Lazo-Cortés, and J. A. Carrasco-Ochoa, "The Impact of Basic Matrix Dimension on the Performance of Algorithms for Computing Typical Testors," in *Lecture Notes in Computer Science*: Springer, 2018.
- [41] V. Rodríguez-Diez, J. F. Martínez-Trinidad, M. S. Lazo-Cortés, and J. A. Carrasco-Ochoa, "A new algorithm for reduct computation based on gap elimination and attribute contribution," *Information Sciences*, vol. 435, pp. 111-123, 2018.
- [42] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.
- [43] Y. Villuendas-Rey, C. F. Rey-Benguría, Á. Ferreira-Santiago, O. Camacho-Nieto, and C. Yáñez-Márquez, "The Naïve Associative Classifier (NAC): A novel, simple, transparent, and accurate classification model evaluated on financial data," *Neurocomputing* vol. 265, pp. 105-115, 2017.
- [44] Y. Villuendas-Rey, M.-D. Alanis-Tamez, C.-F. Rey Benguría, C. Yáñez-Márquez, and O. Camacho-Nieto, "Medical Diagnosis of Chronic Diseases Based on a Novel Computational Intelligence Algorithm," *Journal of Universal Computer Science*, vol. 24, no. 6, pp. 775-796, 2018.
- [45] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1-34, 1997.
- [46] J. A. Hernández-Castaño, O. Camacho-Nieto, Y. Villuendas-Rey, and C. Yáñez Márquez, "Experimental Platform for Intelligent Computing (EPIC)," *Computación y Sistemas*, vol. 22, no. 1, pp. 245-253, 2018.
- [47] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10-18, 2009.
- [48] J. Alcalá-Fdez et al., "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, pp. 255-287, 2011. P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-Validation," in *Encyclopedia of Database Systems*: Springer, 2009.
- [49] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation," in *Lecture Notes in Computer Science*: Springer, 2006.
- [50] M. Kubat, R. Holte, and S. Matwin, "Learning when negative examples abound," in *European Conference on Machine Learning*, 1997.
- [51] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.
- [52] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *International Conference on Machine Learning*, 1997.
- [53] D. Ballabio, F. Grisoni, and R. Todeschini, "Multivariate comparison of classification performance measures," *Chemometrics and Intelligent Laboratory Systems*, vol. 174, pp. 33-44, 2018.

An Empirical Investigation of the Relationship Between Business Process Transparency and Business Process Attack

Alhanouf Aldayel¹, Dr. Ahmad Alturki²
College of Computer and Information Science
King Saud University, Riyadh, Saudi Arabia

Abstract—Business Process Management (BPM) is a management approach to discover, analyze, redesign, execute and monitor business processes. Implementing BPM concepts help and benefit organizations by increasing their productivity, achieving their strategies and operational excellence, and saving costs. Rosemann et al. identify business process transparency as one of the key values of BPM, and essential to achieving other BPM benefits. Business process transparency provides visibility about how operations/activities are conducted in a detailed way, sometimes with technical details, within an organization; which facilitates the identification of structural issues of the process model. A conducted content analysis of the literature shows that fraudsters have leveraged structural issues of the business process model to commit fraud. Such fraud can be labeled as a Business Process Attack (BPA). In analogy to information system security attack, BPA can be defined as the exploitation of a vulnerability in a business process model to commit fraudulent activities that influence the business negatively such as achieving invalid or unwanted results. This research aims to investigate the relationship between the degree of business process transparency and exposure to BPA. If the relationship is positive, appropriate security controls shall be implemented on the business process transparency to avoid BPA. The main research question is: What is the relationship between an organization's degree of business process transparency and exposure to BPA. A quantitative research method is employed to measure and understand the impact of business process transparency on BPA. An experiment is designed and conducted to assess the awareness of the existence of vulnerabilities in various process models and how to exploit them to commit BPA. Results show that there is a positive significant relationship between increased business process transparency and exposure to BPA. This research contributes towards understanding and highlights the negative impact of business process transparency, which motivates researchers to investigate this phenomenon and find appropriate solutions.

Keywords—Business Process Management (BPM); business process; transparency; business process attack; fraud

I. INTRODUCTION

Business Process Management (BPM) is “a body of methods, techniques and tools to discover, analyze, redesign, execute and monitor business processes ” [1,p. 5]. A business process is a set of logically related activities performed to achieve the desired business outcome. Business processes are managed by BPM, which is an essential management guide for organizing and managing business processes using well-known methods, techniques, and tools to manage business processes

[1]. BPM is applied by a defined sequence of activities known as the BPM lifecycle. It consists of six stages: process identification, process discovery, process analysis, process redesign, process implementation, and process monitoring and controlling [1].

The BPM approach is becoming widely adopted, and BPM research has become interested in analyzing the perceived effects of applying such an approach. BPM aims to achieve both strategic and operative organizational goals [2]. It helps organizations in increasing their productivity, achieving operational excellence, and saving costs [3].

Recent research by Rosemann et al. [4] introduced the value-driven BPM framework, which consists of seven values. The first six values are grouped as three pairs of opposing values that alleviate three classical business conflicts (efficiency–quality; agility–compliance; and integration–networking). The seventh value is transparency, which Rosemann et al. consider as the core value of BPM, and provides visibility into an organization's operations.

Transparency in BPM provides visibility regarding how operations are conducted and enhances decision-making processes in organizations[4]. In their work Rosman et. Al.

Mentioned that a process model repository can be published via various channels like an intranet. However, they did not mention publishing to external parties specifically. A study by Kohlbacher et al. [5] shows that higher transparency facilitates the identification of problems in a business process. Because process transparency entails the transparency of process weakness such as structural issues in the process model if there are any.

Structural issues can constitute an opportunity enabling fraudsters to commit fraudulent activities. A literature review shows several fraud cases that originate from the exploitation of different vulnerabilities in process models. For instance, the Swiss bank UBS had a loss of approximately two billion US dollars due to a structural issue of the process model [6]. In Europe, processes that include "forward-settling" exchange-traded funds (ETF) cash options do not issue confirmations until after settlement has taken place. Fraudsters use this vulnerability in the process to receive payment for a trade before the transaction is confirmed. While the cash cannot be simply retrieved, the seller may still show the cash on their books and possibly use it in further transactions. This allowed

for a recursive series of transactions, creating an ever-growing snowball. Such fraud cases can be considered as instances of business process attack (BPA).

In analogy to information system attack which is defined as the act of exploiting a vulnerability in a controlled system to damage or steal an organization's information or physical asset [7], researchers of the current study define BPA as 'the exploitation of a vulnerability in a business process model to commit fraudulent activities that influence the business negatively'. To avoid attacks, organizations need to be aware of situations that lead to attacks which then secures themselves with appropriate security controls.

The current study aims to investigate the relationship between business process transparency and BPA. If the relationship is positive, appropriate security controls shall be implemented on the business process transparency to avoid BPA. The main research question is: What is the relationship between an organization's degree of business process transparency and exposure to BPA.

II. LITERATURE REVIEW

A. BPM Security

BPM security aims to provide sound guarantees regarding adherence to security, privacy, and regulatory compliance requirements. Security must be seamlessly integrated and applied to business processes at every stage of the BPM lifecycle. To achieve BPM security organizations, need to understand where and when security requirements are required. The security extended enterprise meta-model is used for this purpose. The model divides BPM security into three layers: the business layer, the application layer, and the infrastructure layer (Figure 1). The business layer defines the business processes and organizational structure to be followed. The application layer defines the security needed by the services and the data schemas required for the execution of the business processes. The infrastructure layer defines the security needed for the software and hardware to automate the execution of business processes[6].

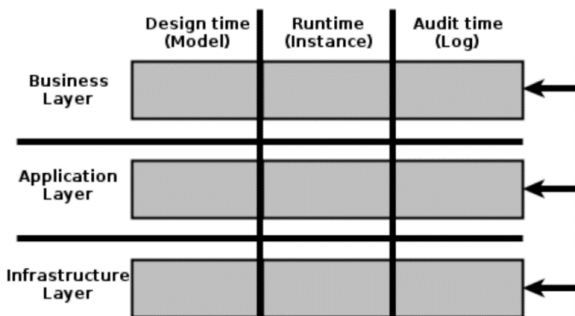


Fig. 1. Security Extended Enterprise Metamodel [6].

Business process security must consider security each of the three layers. Each layer is divided into three stages according to the timepoint and entity where they act. Design time is concerned with process models, the runtime is concerned with process instances, and audit time is concerned with event logs. Both the application layer and infrastructure layer have been heavily investigated; however, the business

layer is a relatively new and challenging area for research. Such research focuses on business process design time security; the security of the process model is investigated before the actual runtime of the business process instances[6].

1) *BPM security requirements*: To ensure the required level of security, organizations shall enforce certain security requirements. Current BPM security requirements focus on design and runtime, and can be classified into the following general types [6, 8]:

- **Need-to-know**: participants should access only the needed sensitive data to execute their tasks.
- **Authorization**: access control is needed to ensure that only authorized roles can execute activities within a process. This requirement is usually achieved with Role-Based Access Control RBAC.
- **Usage control**: to monitor conditions that must hold after the access to a resource e.g., the maximum number of access to a resource this requirement can be used to monitor the compliance to regulatory policies and data protection requirements.
- **Separation of duty**: constraints on process execution are needed to limit the abilities of participants to execute tasks, eventually reducing the risk of fraud e.g., some activities in a process cannot be executed by the same subject or by the same role.
- **Binding of duty**: In contrast, to the separation of duties, this requirement enforces some activities to be executed by the same subject or by the same role. This requirement helps to ensure the integrity of data.
- **Isolation**: Data must stay confidential during the execution of a process.

2) *Gaps in Business Process Management (BPM) Security*: To maintain Business Process management (BPM) security, organizations need to be aware of security threats, and appropriate security controls shall be implemented. Current business process security controls focus on the optimal assignment of subjects, roles, and activities in a Role-Based Access Control setting. In RBAC, each subject acting in a role should only have the minimal permissions necessary to execute the process, and all the assignments that lead to more rights should be prohibited. Such control is designed to prevent Business Process Attacks (BPA) by checking precisely defined security requirements during process execution[9, 10]. However, BPA can occur following the Standard Operating Procedures (SOP) and without any violation of security requirements. The attacks happen using existing vulnerability the business process model structure and become possible by only viewing the business process model. Most organizations do not pay attention to the secure sharing of business process models. And publish them through organizations' intranet. The models are published in an understandable and intuitive format to ensure that business processes are well accepted by the users[10]. organizations

shall pay more attention and only share process models to the intended audience, to reduce the possibility of exposure to BPA.

B. Transparency

The literature shows that there are many definitions of transparency. Some researchers define transparency using a descriptive approach, while others are using a normative approach. Oliver defines transparency using the descriptive approach using three elements: an observer, the object to be observed, and a way of observation. Moser takes a normative approach to define transparency: "to open up the working procedures not immediately visible to those not directly involved to demonstrate the good working of an institution" [11, p.258]. The normative approach does not only describe what transparency is but also what is needed for it to be achieved [11].

The variation in the definitions of transparency does not only come from the definition type but also from the context in which it is being used. For example, In the context of strategic alliances, Ackerman et al. define transparency as "sharing data regarding current order and production statuses as well as plans and forecasts with various supply chain partners" [12, p.4]. In the context of financial markets, Madhavan et al. define it as the "ability of market participants to observe information about the trading process" [12, p.4]. In the context of organizational governance, Potosky defines transparency as the "extent to which a communication medium facilitates a clear or unobstructed communication exchange" [12, p.4]. In the context of the electronic market, it is defined as the "degree of visibility and accessibility of information" [12, p.4].

There are some efforts to generalize the definition of transparency. For example, Davis defines transparency as "lifting the veil of secrecy" [11, p.258]. Hood defines it as "openness to public scrutiny" [13, p.5]. Such definitions are typically broader in scope; however, they do not specifically indicate all the elements of transparency. Schnackenberg et al. [12] have studied transparency definitions through the literature, concluding with a general definition of transparency: "Transparency is the perceived quality of intentionally shared information from a sender" [12, p.5]. Based on this definition, they suggest a conceptualization of transparency by examining the quality of information using three primary manners: disclosure, clarity, and accuracy.

Disclosure means: "the perception that relevant information is received in a timely manner" [12, p.9]. Disclosure implies that information must be openly shared for it to be considered transparent. Researchers see disclosure as a central dimension of transparency. Pirson et al., [12] for example, measure transparency as a stakeholder's perception that firms openly share all relevant information. Perotti et al. [12] suggest that perceptions of transparency are built around a stakeholder's ability to gather the necessary information about a firm. Williams [12] describes disclosure in four processes: analysis, interpretation, documentation, and communication—in analysis, the target audience is identified; in interpretation, the relevant information for the audiences is determined; in documentation, the relevant information is documented; and in communication, information is distributed to internal and

external audiences. Documentation and communication are associated with the open release of information, while analysis and interpretation are important to distinguish relevant from irrelevant information [12].

Clarity means "the perceived level of lucidity and comprehensibility of information received from a sender" [12, p.9]. The information must be understandable to be considered transparent. Complicated mathematical algorithms cannot be considered transparent even if highly disclosed. Daft and Lengel find that a major problem for transparency is a lack of informational clarity rather than a lack of data sharing (disclosure) [12]. Rawlins [14] argues that transparency is not only achieved by disclosure but also by increasing understandability. In this statement, Gower highlights that transparency implies an increase in the understanding of parties who are interested in the actions or decisions of an organization [14].

Accuracy means "the perception that information is correct to the extent possible given the relationship between sender and receiver" [12, p.10]. Information cannot be considered transparent if it is biased or incorrect. Bushman et al. suggest that information must be valid for it to be considered transparent [12].

1) *Organizational transparency*: Nowadays, transparency is an unambiguously positive concept. Without transparency, the actions of organizations cannot be monitored. To ensure organizations comply with the law and public interest, organizations need to be transparent [15]. Higher organizational transparency improves the image of an organization in the global market and toward the public [1]. Organizations benefit from organizational transparency by improved organizational efficiency and the effectiveness of the decision-making process [14]. It also plays an important role in facilitating business globalization. Transparency provides customers with the confidence they need when dealing with foreign companies that obey other countries' laws. Moreover, providing information disclosure has a positive relationship with organizational performance. And is an enabler for observability, accountability, certainty, and better conduct [16].

Transparency is categorized into many types based on different perspectives. Based on the type of information in question, Heald divides transparency into event transparency and process transparency. Event transparency provides information about what organizations achieve in the form of inputs, outputs, and outcomes (i.e., organizational performance reports). Process transparency provides information about how organizations achieve this outcome (i.e. governmental transformation process) [17].

Bannister et al. [13] proposed a modified version of Heald's model to adapt it to computer-mediated transparency. The new E-transparency model consists of three categories: data transparency, process transparency, and decision/policy transparency. Similar to event transparency, data transparency is concerned with what organizations are doing. Facts and figures are used to provide data transparency. Process

transparency is concerned with how organizations are working. The steps of organizational processes should be clarified for them to make the process transparent. Decision/policy transparency is concerned with why organizations are doing their work in a specific way. An organization must justify its decisions to be decision transparent.

Heald also categorizes transparency according to its direction—upwards/downwards (a vertical dimension), and inwards/outwards (a horizontal dimension). The vertical dimension transparency goes through organizational hierarchy directions. In the downward direction, managers will be able to monitor their employees' actions. If the relation is symmetric, the upward direction will allow employees to view their managers' actions. The vertical dimension addresses an organization's internal transparency, whereas the horizontal dimension addresses an organization's external transparency. In an inward direction, an organization's internal actions can be seen from the outside. In an outward direction, the organization can see external actions. An organization is said to have full symmetric transparency when all dimensions are present at the same time [18].

2) *BPM transparency*: A key value of adopting Business process management (BPM) is providing an organization with business process transparency which provides visibility about how operations/activities are conducted in a detailed way[4]. Business process transparency is essential to achieve other BPM values such as agility, quality, networking, integration, efficiency, and compliance[4]. The use of computer systems to manage business processes empowers business process transparency and allows organizations to achieve high organizational transparency i.e. data transparency, process transparency, and decision/policy transparency [4].

3) *Business Process Model Abstraction (BPMA)*: Business Process Model Abstraction (BPMA) is a technique applied to detailed process models to produce generalized versions of the process model[19]. Two main methods are used to apply BPMA: elimination and aggregation. Elimination omits some process model elements of the detailed version to generate the abstracted version. While aggregation groups related process elements of the detailed version to generate an abstracted version. Both methods hide certain activities of the abstracted version and hence reduce process transparency. BPMA shall assure that the resulted abstract process model is well-formed and maintains the original process semantics[20].

BPMA is conducted by applying a set of atomic abstractions are on the initial detailed model. An abstraction is a function that takes a process model as an input and produces a process model as an output. Based on selected criteria, each abstraction hides some process details and brings the model to a higher degree of abstraction. individual abstractions can be combined and afterward controlled to deliver the desired abstraction level[20]. Selecting abstraction criteria can be based on roles activity frequency or activity completion time, or structural aspects of a process model[19].

Moreover, abstraction criteria can be based on functional aspects such as sequential, block, and loop abstractions. Sequential abstraction replaces a sequence of tasks and events by one aggregated function[20]. In Block abstraction, a process fragment in the model enclosed between connectors is replaced with one function. The replaced fragment usually represents parallelism or a decision point in a process. In loop abstraction, Iterated tasks are replaced with a loop construct iterated for successful process completion. In a process model, the fragment to be repeated is enclosed into a loop construct.

C. Fraud

Fraud is defined as the art of deception for gain. Fraud is always intentional. According to Brenner, when someone commits fraud, four elements are present: the perpetrator communicates false statements to the victim, the perpetrator communicates what they know are false statements with the intent of defrauding the victim, the perpetrator's statements are false, and the victim is defrauded out of something of value [21].

According to the fraud triangle theory, fraud occurs in a situation where three components are present: opportunity, pressure, and rationalization. Opportunity refers to the opportunity for the perpetrator to commit fraud (i.e. the lack of internal controls creates an opportunity). Pressure refers to the motivation or driving force behind committing fraud (i.e. personal financial need could cause pressure). Rationalization refers to the fraudsters' justifications for the fraudulent activity using cognitive reasoning. Fraudsters rationalize fraud to consider their act acceptable [22].

1) *Organizational fraud*: In organizations, fraud can be categorized into two main categories. The first category is fraud committed by organizations regarding their financial reporting—i.e., when they use false financial reports to intentionally defraud investors and third parties to benefit the organization. An example of this category is financial statement fraud. Here, organizations intentionally misstate figures and make false disclosures in financial reports to deceive financial statement clients.

The second category is the fraud perpetrated against an organization that results in harm to the organization itself. An example of this category is employee fraud. This fraud includes the theft of cash or inventory, skimming revenues, and payroll fraud [23]. The fraud against the organization can be committed either internally by employees, or externally by someone who's externally related to the organizations such as suppliers, and other parties [24]. Both profit and non-profit organizations are susceptible to both categories of fraud [25].

2) *BPM fraud*: Organizations adopting a BPM approach are not excluded from being susceptible to organizational fraud. Fraud related to the business process is known as process-based fraud (PBF) and is enabled by deviations from standard operating procedures (SOP). It can be detected by analyzing deviations in throughput time such as duty sequences, wrong duty decisions, or wrong duty combinations. Process execution information is usually stored in an event log. This information includes events, originators, and time

stamps. Control flow analysis can be used to analyze process information patterns from event logs. Cases, where the fitness function is small, are considered as noise. This noise is identified as suspicious PBF.

Process-based fraud causes deviations from the process model. However, fraud cases can occur even during the normal flow of running processes. An example is a case of fraud in one of the leading finance companies in Sri Lanka. A fraud case was detected during an audit check. It was found that several returned checks for different clients in a specific branch were issued from the same checkbook owned by a marketing officer working in the branch. The marketing officer is responsible for initiating the contracts of returned checks. The marketing officer created personal agreements with clients whereby the client would pay in cash in advance to get a discount, and the marketing officer would subsequently invest the money for personal benefit and earn a return. In this case, the fraud did not cause the organization any financial loss; however, it could damage the reputation of the organization [26]. Additionally, the Swiss bank UBS had a loss of approximately two billion US dollars due to the use of "forward-settling" ETF cash positions [6].

In Europe, processes that include ETF do not issue confirmations until after settlement has taken place. This vulnerability in the process model can be exploited by a party to receive payment for a trade before the transaction is confirmed. While the cash cannot be simply retrieved, the seller may still show the cash on their books and possibly use it in further transactions. This will allow for a recursive series of transactions, creating an ever-growing snowball.

In both cases, the fraudster exploited a vulnerability in the process model to commit the fraud. In the case of the finance company, an absence of internal controls and policy in the case of checks returned enabled the fraudster to commit the fraud, while in the Swiss bank UBS, weak process design allowed the fraud to be committed. Eventually, the fraud in both cases affected the organizations negatively.

More fraud cases can happen without deviation in SOP, a case of a man in China clearly explain how weakness in the process structure enables fraudsters to commit such fraud. The man purchased one First class airline ticket, and used it to have a year of free meals! The man just used his ticket as a regular traveler to have a meal in the first-class lounge; however, instead of getting in the flight, he kept rescheduling his flight to another day. The man will show up on the rescheduled date in the lounge with a newly issued ticket, eat his meal, and reschedule his flight again! Airlines staff discovered that the man rescheduled the same ticket over three hundred times in a year. Moreover, when the airlines started investigating the fraud, the man simply canceled his ticket before the expiration date and had a full refund [27]. Such fraud cases involve more risk as they are harder to detect by organizations.

It is important to define a broader scope of process fraud that combines fraud that causes deviation in the SOP, as in PBF cases, and those that occur without causing such deviation. In both cases, the action of exploiting vulnerabilities to commit fraud is similar to an information system security

attack. An attack is a deliberate act to exploit a vulnerability in a controlled system to damage or steal an organization's information or physical asset. Both process fraud and security attacks exploit a vulnerability and affect organizations negatively; however, the entities that are vulnerable to exploitation differ. In the case of fraud, the targeted entity is the business process, whereas, in an information security attack, it is the system. The author uses the term business process attack (BPA) to describe this act.

D. Transparency and Fraud

Transparency is nowadays an unambiguously positive concept. It ensures an organization's compliance with the law and the public interest. Higher transparency improves the image of the organization in the global market and toward the public [28]. Transparency helps organizations to improve the efficiency and effectiveness of their decision-making process [14]. It also plays an important role in facilitating business globalization by providing customers with the confidence they need when dealing with foreign companies that obey other countries' laws. Most of the published research within organizational transparency focus on its positive effects.

Rosemann et al. consider business process transparency as the core value of BPM, which provides visibility into an organization's operations [4]. Previous research shows that higher transparency facilitates the identification of problems in business processes to organizations. Such Identification help organization to optimize the weakened business processes. A study by Kohlbacher et al. [29] included 44 process-oriented firms; the results showed that process orientation leads to higher transparency, which enhances the identification of organizational problems and their causes. Malinova et al. [30] studied reasons for BPM adoption, finding that considerable numbers of organizations adopted the BPM approach mainly for identifying process weaknesses; they argue that without BPM practices, this would be more difficult or even impossible [29]. However easy identification of process weaknesses can facilitate fraud. According to Wells et al. [31], fraud is commonly committed by people who know the weaknesses and how to exploit them best.

Just as in the case of software programming, source code transparency in Open-Source software gives both attackers and defenders the analytic power to do something about known source code vulnerabilities, however, If the defender didn't improve security or eliminate vulnerabilities, Attackers will be able to use them in malicious attacks [32]. In the same way, process transparency provides visibility to process weaknesses which constitute vulnerabilities that can be used to commit fraudulent activities. If organizations did nothing about discovered vulnerabilities, a fraud opportunity exists and is available to fraudsters. According to the fraud triangle theory, fraud occurs in a situation where three components are present: opportunity, pressure, and rationalization [22].

E. Related Work

Wehmeier et al. analyzed several published research studies on transparency, and find that more than half of them focus on its positive impact [14]. Research calls upon organizational transparency focused on its relationship with trust. Researchers find that increased transparency increase employees' trust in

their organizations [33]. Moreover, it helps in creating, maintaining, and repairing confidence and trust in an organization- stakeholder relationships [12]. In their work, Vössing et al. [34] find that organizations can enhance process performance by making process information accessible to their employees.

In the contrast to the widespread belief about its benefits, however, transparency is indeed a double-edged sword [35]. Fewer research studies the negative impact of transparency [14], research shows that increased transparency in buyer-supplier relationships may cause buyer's frustration. In the electronic marketplace, the increased transparency can harden the creation of a close buyer relationship. Because it facilitates the comparison of organizations' products with other competitor's products and causes organizations' products to be commoditized [35]. In the health care field [36], increased process transparency results in a decrease in trust in the health care system. Trust levels were higher among the group given no information about the procedures.

Based on the conducted content analysis of the literature, few number studies focused on the negative impact of increased transparency in BPM; there is a lack of empirical research studies to investigate the relationship between business process transparencies to Business Process attack (BPA). Research on BPM transparency has only discussed its positive impact in terms of decision making and enhancing business process models by understanding business process weaknesses [4, 5]. However, previous research has study the factors which increase the possibility of exposure to a. Process-based Fraud (PBF) - an instance of BPA.

Some characteristics of business process model design have been linked to the possibility of exposure to PBF. Possibility of exposure to PBF increases in business process models that were designed to allow for the skip of some task execution, or the skip of decision and proceed to the next task execution [37]. Moreover, PBF is more likely to occur when a process is allowed to be executed by an unauthorized resource, or when different authorities are given to the same originator. Moreover, the possibility of exposure to PBF has been linked to personal perpetrator behavior. Research findings show that PBF is more likely to occur by perpetrators who are known for

their bad behaviors [37]. This research studies process transparency as a factor that may increase the possibility of exposure to a BPA.

III. HYPOTHESIS

An opportunity for BPA exists when vulnerabilities such as weak security controls are implemented in a business process. Low process transparency can hide the existence of such vulnerabilities because it provides fewer process details. Moreover, low process transparency limits people's understanding of process details, and hence increases people's trust in business processes [36]. This is because they assume that organizations are implementing high standards, even when they are not doing so [36]. Such an attitude may make people unaware of BPA opportunities because they are assuming high-security controls are implemented. To test this assumption, we need to assess people's understandability of a BPA opportunity, in relation to different levels of business process transparency. Two hypotheses are proposed:

H0: Increased business process transparency does not increase attackers' understandability of a BPA opportunity.

H1: Increased business process transparency increases attackers' understandability of a BPA opportunity.

IV. EXPERIMENT DESIGN

For the aim of this study, a single factor experiment is suitable, because it allows investigating the effects of one factor on a common response variable. It also allows analyzing variations of a factor, the factor levels. The response variable, is then, determined by the participants, subjects, in relation to a specific factor level applied to a particular object [38].

The experiment in the current research is designed similarly to the one used in [38] to assess modularity's impact on process understanding. In the experiment, variations of a factor (process transparency degree): The factor levels (high-low) are analyzed. The response variable (level of BPA opportunity understanding) is determined by the participants in the experiment when they interact with different factor levels applied to a particular object (process model). The overall design of the experiment is depicted in Figure 2.

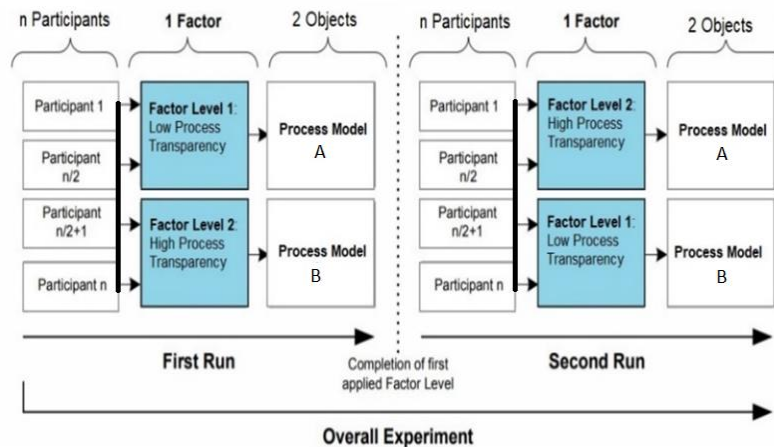


Fig. 2. Experiment Design [38].

V. EXPERIMENTAL SETUP

A. Subjects

The subjects are the people who participate in this experiment. The participants are randomly assigned into two runs (previously explained in experimental design). In the first run, half the subjects will be shown two models: high transparency purchase-to-pay process model and low transparency attend an event model. In the second run, the other half will be shown two different models low transparency purchase-to-pay process model and the high transparency Attend an event model. This way each participant will receive the two different processes (purchase-to-pay and attend an event) in two different transparency degrees (high transparency and low transparency).

B. Objects

The objects to be used in the experiment are different business process models designed with various structural issues. For each process model, two versions are designed with different process transparency levels. The first version is designed with low process transparency and shall include minimal details to understand the business process model. The second version shall be higher in business process transparency, and shall be modeled in detail to make the models more transparent and understandable. The business Process Model Abstraction (BPMA) technique is utilized to generate low transparency versions of the process model [19]. BPMA assures that the resulted process model is well-formed and maintains the original process semantics [20].

Two business processes are selected: purchase-to-pay, and Attend an event, because they are commonly susceptible to fraud. The original process models were re-designed to contain a vulnerability, which represents a common fraud. Figure 3 shows the vulnerable purchase-to-pay business process model. This process allows the staff of a company to request the purchase of goods needed by the company. The process is vulnerable to BPA because no internal controls exist to prevent billing schemes and check tampering. This allows the procurement officer to create fraudulent purchases of goods or services that do not exist, are overpriced, or unnecessary.

Figure 4 shows the vulnerable attend an event process model. This process allows people to buy tickets to attend a specific event. The process is vulnerable to BPA because no checks are done on the attendee's age before entering the event. People over 18 years can illegally use a child's ticket to enter the event. The event organizer will be affected financially and lose money.

C. Factor and Factor Levels

The process transparency degree is the considered factor, with factor levels 'high' and 'low'.

D. Response Variable

The response variable in this experiment is the level of BPA opportunity understanding that the respondents show with respect to the process models, both in their high-transparency and low-transparency versions. To measure the response variable, a specific set of questions are developed for each

version to be answered by the participants. The percentage of correctly answered questions by a subject is used as a measure for the participant's level of understanding of BPA opportunity within a particular model. This approach is previously applied in studies to measure process model understandability [39-41].

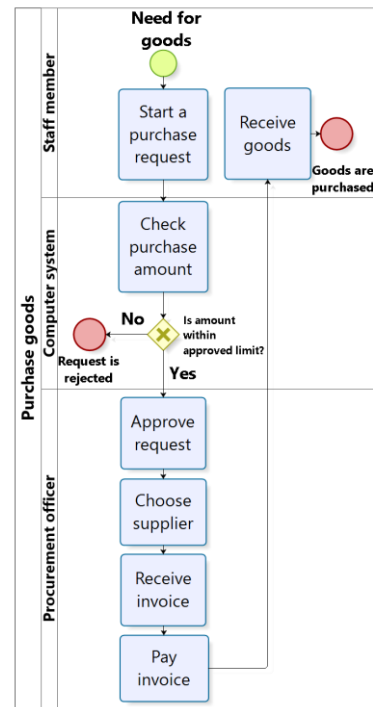


Fig. 3. Vulnerable Purchase-to-pay Process Model (High Transparency).

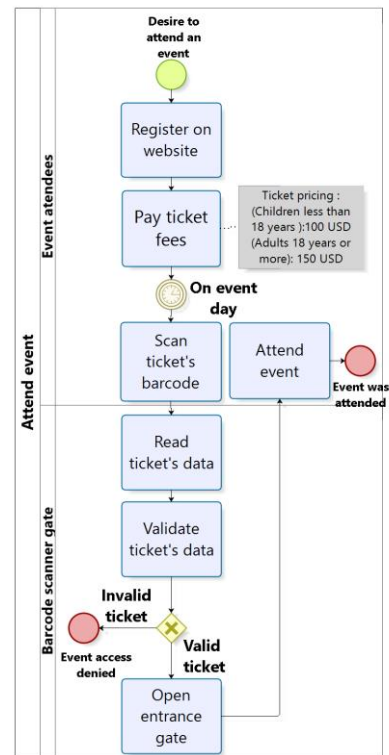


Fig. 4. Vulnerable Attend an Event Process Model (High Transparency).

E. Instrumentation

The experiment was then carried out using a surveying website. The participants use their personal computers or mobile phones to view the process models and answer the questions - about BPA opportunity- under each model. The researcher makes sure that the display of the process model fits both personal computer displays and mobile displays. Also, a note is added under each model to guide the participant on how to enlarge models by zooming – in case the model display is unreadable by the participant.

F. Data Collection Method

Various data collection methods are used to conduct research including interviews, experiments, and questionnaires [42]. The current study uses a structured questionnaire as a method for data collection utilizing participants' responses to a structured set of questions [42]. Questions in a questionnaire can be open or closed. Open questions allow respondents to answer on their own [43]. Open questions are used when a researcher cannot predict what the responses might be [43]. On the other hand, closed questions allow respondents to choose an answer from a set of alternative answers, which makes answers to be more objective [43]. Questionnaires can be conducted in two ways: interview-based, and self-completed. This research makes use of a web-based self-completed questionnaire because, to reach a large sample, and maintain respondents' confidentiality [42]. The questions are going to be closed questions for more objective responses of respondent's awareness of exposure to BPA.

G. Questionnaire Design

There are two questionnaires used in this research- one for each run. Both questionnaires have the same structure, however, some model-specific questions may differ based on the process models selected in each run.

Each questionnaire is structured as follows: the first part contains a question on different business process models to assess respondent's knowledge about exposure to Business Process Attack (BPA). The second part will collect data about participants' previous experience in the field of BPM.

H. Questionnaire Validity and Reliability

To ensure the validity of the research method, the questionnaire used is built based on knowledge acquired by the researcher in the literature review. The questionnaire is also reviewed by the researcher's supervisor. And a pilot test has been undertaken. Moreover, guidelines of questionnaire design are taken into account to assure its validity. Five answers are provided for each question to reduce the chance of answering correctly by coincidence. Answers are positively formulated answers because negations distract respondent's attention.

The reliability of collected data is dependent on the integrity of provided answers. To encourage the respondents' honesty and integrity, the researcher uses a simple design and clear structured questionnaire and insures respondent's confidentiality. Moreover, filter questions are added to each model. A filter question is a question on some aspects of the model and is used to make sure the participant had read the model before answering the questions written based on the

model. The questionnaire considers only the answers of the participants who answered filter questions correctly, which limits the chance for randomly answering the questionnaire.

I. Data Evaluation

After has been collected, the next step done is data evaluation. It involves tasks editing and coding. Editing is used to ensure that questionnaire results are checked for any potential errors or inconsistencies. Coding is used to define the values of different sets of responses. Coding is important to transform questionnaire results into a format that can be easily fed to analytical tools [27]. Correct Answers are coded as "1", and wrong answers are coded as "0".

J. Sampling Method

A sample is a subset of the population. A well-defined sample should have the same characteristics as the population, if not, then the research results will be wrong [27]. In our study, because Business Process Attack (BPA) assumes that attack can happen by anyone who interacts with process models. The target population can include all people, and the selected sample should represent people from different demographics. However, the participants' experience in process modeling can influence the questionnaire results [44]. Because participants' integrate their previous experience with process model content to construct new knowledge [45], which gives an advantage of an experienced user to gain more knowledge about a process model. To avoid such influence on participant experience, this variable should be randomized.

Another important consideration of sampling is to determine the sample size. The bigger the sample is the greater will be its accuracy [27]. Hair et al. [46], suggest that the minimum sample size is five respondents per variable to be analyzed. In this research, there are two variables in each run (high transparency, low transparency) which makes the minimum participants in each run 10 participants, and 20 participants for the overall experiment.

In this research, the participants will be 200 people randomly assigned into the two experiment runs (previously explained in experimental design). In the first run, half the subjects (100 people) were shown two models: high transparency purchase-to-pay process model and low transparency Attend event model. The great number of participants will increase the accuracy of the experiment results.

K. Data Analysis Method

Data analysis is the interpretation of collected data using different analytical tools. According to the requirements of the management is called analysis. Several tools are used for statistical analysis (such as SPSS and Microsoft Excel). The research makes use T-test to assess the significance of the difference between participants' knowledge in each run of the experiment. The result is demonstrated and interpreted based on knowledge acquired by the researchers [27].

VI. RESULTS

To distill the experiment results, a comparison is conducted between participants' performance for each model in terms of the number of correct answers. This helps us to understand if

more transparency will increase people's understandability of exposure to BPA. The percentage of correct answers for each model variant is calculated; as shown in table 1; the result for the high transparency version is analysed in comparison to the low transparency version for each process model.

TABLE I. AVERAGE PERCENTAGES OF CORRECT ANSWERS FOR EACH MODEL VARIANT

Process model / Transparency degree	Low transparency	High transparency
"Purchase-to-Pay"	38%	44%
"Attend an Event"	43%	51%

A. Purchase-to-Pay Process Model Results

Looking at the results reported in table-1, we can see that when participants are given the low transparency version of the "purchase-to-pay" process model they had correctly answered 38% of the questions about BPA opportunity. This percentage increases to 44% for the participant who had been given the high transparency version of the "purchase-to-pay" process model. To test if this increase is statically significant, a T-test is used.

To use the T-test, two assumptions must be met: data should be normally distributed, and the two samples should have equal variance. First, we check if data is normally distributed. Data is normally distributed when the standardized skewness and standardized kurtosis should be within the range of -2 to +2 for each model variant. For the "purchase-to-pay" process model, the actual results of skewness are (1.24, .89), and the results of kurtosis are (-.84, -1.24) for the low transparency model version, and the high transparency model version respectively. Since all values of skewness and kurtosis are within the range of -2 to +2, we can assume that percentage of correct answers for the "purchase-to-pay" process model is normally distributed. Second, the two samples should be tested using F-tests to ensure they have equal variance. Applying F-test with 95% confidence shows that standard deviations of the samples for each of the models are the same. Table 2 shows the results of applying of F-test on purchase-to-pay low transparency business process model and high transparency business process model version.

TABLE II. F-TEST (PURCHASE-TO -PAY BUSINESS PROCESS MODEL)

	Purchase-to-pay high transparency	Purchase-to-pay low transparency
Mean	0.44	0.38
Variance	0.08	0.08
Observations	100.00	100.00
Df	99.00	99.00
F	1.03	
P(F<=f) one-tail	0.44	
F Critical one-tail	1.39	

As T-test assumptions are met for the "purchase-to-pay" process model, we can apply T-test results which generate a P-value for the comparison between process model variants. A P-value lower than 0.05 is considered significant. The T-test results with (P=.13). P-value suggests there is no difference between the high transparency version of the process model and the low transparency version in terms of the average percentage of correctly answered questions on exposure to BPA. We can conclude that the increase in the number of correct answers between the low transparency version of the "purchase-to-pay" process model and the higher transparency one is statically significant.

B. Attend an Event Process Model Results

Looking at the results reported in table-1 above, we can see that when participants are given the low transparency version of the "attend an event" process model they had correctly answered 43% of the questions about BPA opportunity. This percentage increases to 51% for the participant who had been given the high transparency version of the "attend an event" process model. To test if this increase is statically significant, a T-test will be used.

To use the T-test, two assumptions must be met: data should be normally distributed, and the two samples should have equal variance. First, we check if data is normally distributed. Data is normally distributed when the standardized skewness and standardized kurtosis should be within the range of -2 to +2 for each model variant. For the "attend an event" process model, the actual results of skewness are (.45,.1), and the results of kurtosis are (-.48, -1.11) for the low transparency model version, and the high transparency model version respectively. Since all values of skewness and kurtosis are within the range of -2 to +2, we can assume that percentage of correct answers for the "attend an event" process model is normally distributed. Second, the two samples should be tested using F-tests to ensure they have equal variance. Applying F-test with 95% confidence shows that standard deviations of the samples for each of the models are the same. Table 3 shows the results of applying of F-test on the "attend an event" low transparency business process model and high transparency business process model version.

TABLE III. F-TEST (ATTEND AN EVENT BUSINESS PROCESS MODEL)

	Attend an event high transparency	Attend an event low transparency
Mean	0.51	0.43
Variance	0.10	0.08
Observations	100.00	100.00
Df	99.00	99.00
F	1.28	
P(F<=f) one-tail	0.11	
F Critical one-tail	1.39	

As T-test assumptions are met for the "attend an event" process model, we can apply T-test results which generate a P-value for the comparison between process model variants. A P-value lower than 0.05 is considered significant. The T-test results with $(P=.06)$. P-value suggests there is no difference between the high transparency version of the process model and the low transparency version in terms of the average percentage of correctly answered questions on exposure to BPA. We can conclude that the increase in the number of correct answers between the low transparency version of the "attend an event" process model and the higher transparency one is statically significant.

VII. TESTING HYPOTHESIS

A. *H0: Increased Business Process Transparency does not Increase Attackers' understandability of a BPA Opportunity*

To test this hypothesis, we need to assess the P values in regard to the two process models used in the experiment. If the P-value is significant ($\leq .05$) then H0 will be accepted, otherwise, when the P-value is greater than (.05) the H0 will be rejected. Looking at the P values reported in the previous section (5.3), P values are (.13,.06) both "purchase-to-pay" model and "attend an event" process model respectively. we can conclude that H0 is rejected.

B. *H1: Increased Business Process Transparency Increases Attackers' understandability of a BPA Opportunity*

This hypothesis is the alternative hypothesis of H0 "Increased business process transparency does not increase attackers' understandability of a BPA opportunity." the H1 is accepted when H0 is rejected in vise versa. Since H0 is rejected, we can conclude that H1 is accepted. And we can say that increased business process transparency does increase attackers' understandability of a BPA opportunity.

VIII. DISCUSSION

In general transparency in BPM is a positive value. Transparency is categorized according to its direction. The vertical dimension addresses an organization's internal transparency, whereas the horizontal dimension addresses an organization's external transparency.

Internal process transparency provides visibility about how operations/activities are conducted in a detailed way[4]. Internal Process transparency is essential to achieve other BPM values such as agility, quality, networking, integration, efficiency, and compliance[4]. External process transparency can improve customer relationships. However, transparency is indeed a double-edged sword [35].

Studies show that higher transparency facilitates the identification of problems in a business process. Such identification of process weaknesses can facilitate fraud. According to Wells et al. [31], fraud is commonly committed by people who know the weaknesses and how to exploit them best.

Just as in the case of software programming, source code transparency in Open-Source software gives both attackers and defenders the analytic power to do something about known

source code vulnerabilities, however, If the defender didn't improve security or eliminate vulnerabilities, Attackers will be able to use them in malicious attacks [32]. In the same way, process transparency provides visibility to process weaknesses which constitute vulnerabilities that can be used to commit fraudulent activities. If organizations did nothing about discovered vulnerabilities, a fraud opportunity exists and is available to fraudsters.

This research aims to investigate the relationship between the degree of business process transparency and exposure to BPA. A single factor experiment is implemented to assess modularity's impact on process understanding. It also allows analyzing variations of a factor (process transparency degree): The factor levels (high-low). Where the response variable (level of BPA opportunity understanding) is determined by the participants in the experiment when they interact with different factor levels applied to a particular object (process model).

Findings suggest that increased business process transparency can constitute an opportunity to commit fraud. The opportunity exists when people understand different vulnerabilities in process models, and who to exploit them the best to commit fraud.

To avoid attacks, organizations need to be aware of situations lead to attack to secure themselves with appropriate security controls.

IX. CONCLUSION

The research main question is: "What is the relationship between an organization's degree of business process transparency and exposure to BPA?". To Answer the research main question, an experiment was conducted using two process models with different variants "high transparency" and "low transparency. Results show that the high transparency of a process model increases participants' understandability of exposure to BPA. And hence increases the risk of being attacked by BPA.

To achieve the benefits of BPM transparency while avoiding the risk of being attacked by BPA, the researcher recommends the following: Organizations need to follow the BPM security model described in the literature review section and ensure that all their processes are free of structural process vulnerabilities during business process design time. Additionally, suspicious process executions should be detected and prevented by using runtime controls and analyzing event logs. Organizations shall enforce well-known BPM security requirements such as (need-to-know, authorization, usage control, separation of duty, and Isolation when needed. BPM security requirements cover regulatory requirements and privacy and data protection requirements. By doing so the organization reduce vulnerabilities in the process model and hence lowers the risk of being susceptible to BPA. Organizations shall also consider business process model privacy and only share process model design to its intended audience. Moreover, the process model shall be saved in a secure location and not shared or saved out of the organization. In certain cases, when it is needed to share the process models with external parties, process models should be considered as

confidential data and Non-Disclosure Agreements (NDA) shall be used.

X. LIMITATIONS

There are some limitations to this research project. The first limitation is in the number of business processes models used during the experiment. Only four process models were considered. The researcher kept the number small to encourage participants to complete the survey and make sure all processes are understood by the participants. For the same reason, the chosen processes were simple, common, business process models. The second limitation is in the research sample. The population was not limited to a specific type or specific characteristics for attackers. There could be other factors that affect the results, however to our best of knowledge, no prior research identifies a special characteristic of business process attackers and anyone can attack a business process.

XI. FUTURE WORK

Future research can be conducted to assess the effect of attackers' experience in BPM and business process model understandability on transparent business process model exposure to BPA. Moreover, In regards to the process mining research area, anomaly detection is not very frequently researched [47]. Future research can focus on using Business Intelligence (BI) for vulnerability detections during design time.

REFERENCES

- [1] Dumas, M., *Fundamentals of business process management*. 1st ed. 2013, New York: Springer.
- [2] Meerkamm, S., *The Concept of Process Management in Theory and Practice – A Qualitative Analysis*. 2010: Springer.
- [3] Van Der Aalst, W.M., *Business process management: a comprehensive survey*. ISRN Software Engineering, 2013. 2013.
- [4] Franz, P., M. Kirchmer, and M. Rosemann, *Value-driven business process management—impact and benefits*. 2011.
- [5] Kohlbacher, M. *The perceived effects of business process management in Science and Technology for Humanity (TIC-STH)*, 2009 IEEE Toronto International Conference. 2009. IEEE.
- [6] Müller, G. and R. Accorsi, *Why are business processes not secure?*, in *Number Theory and Cryptography*. 2013, Springer. p. 240-254.
- [7] Whitman, M.E. and H.J. Mattord, *Principles of information security*. 2011: Cengage Learning.
- [8] Arsac, W., et al. *Security validation tool for business processes*. in *Proceedings of the 16th ACM symposium on Access control models and technologies*. 2011. ACM.
- [9] Gritzalis, D., et al. *Insider threat: enhancing BPM through social media*. in *New Technologies, Mobility and Security (NTMS)*, 2014 6th International Conference on. 2014. IEEE.
- [10] Brucker, A.D., et al. *SecureBPMN: Modeling and enforcing access control requirements in business processes*. in *Proceedings of the 17th ACM symposium on Access Control Models and Technologies*. 2012. ACM.
- [11] Meijer, A., *Understanding modern transparency*. *International Review of Administrative Sciences*, 2009. 75(2): p. 255-269.
- [12] Schnackenberg, A.K. and E.C. Tomlinson, *Organizational transparency: A new perspective on managing trust in organization-stakeholder relationships*. *Journal of Management*, 2016. 42(7): p. 1784-1810.
- [13] Bannister, F. and R. Connolly, *The trouble with transparency: a critical review of openness in e-government*. *Policy & Internet*, 2011. 3(1): p. 1-30.
- [14] Wehmeier, S. and O. Raaz, *Transparency matters: The concept of organizational transparency in the academic discourse*. *Public Relations Inquiry*, 2012. 1(3): p. 337-366.
- [15] Menéndez-Viso, A., *Black and white transparency: Contradictions of a moral metaphor*. *Ethics and information technology*, 2009. 11(2): p. 155-162.
- [16] Ibini, E. and T. Izims, *Effect of Organizational Transparency on Organizational Performance: A Survey of Insurance Companies in Lagos State Nigeria*. *Journal of Economics, Management and Trade*, 2020: p. 52-62.
- [17] Meijer, A., *Understanding the complex dynamics of transparency*. *Public Administration Review*, 2013. 73(3): p. 429-439.
- [18] Heald, D., *Why is transparency about public expenditure so elusive?* *International Review of Administrative Sciences*, 2012. 78(1): p. 30-49.
- [19] Milani, F., et al., *Criteria and heuristics for business process model decomposition*. *Business & Information Systems Engineering*, 2016. 58(1): p. 7-17.
- [20] Polyvyanyy, A., S. Smirnov, and M. Weske, *Business process model abstraction*, in *Handbook on Business Process Management 1*. 2015, Springer. p. 147-165.
- [21] Vaisu, L., M. Warren, and D. Mackay, *Defining fraud: issues for organizations from an information systems perspective*. *PACIS 2003 Proceedings*, 2003: p. 66.
- [22] Shao, S., *What are Some Best Practices for Internal Controls to Prevent Occupational Fraud in Small Businesses?* 2016.
- [23] Golden, T.W., et al., *A guide to forensic accounting investigation*. 2011: John Wiley & Sons.
- [24] Jans, M., et al., *A business process mining application for internal transaction fraud mitigation*. *Expert Systems with Applications*, 2011. 38(10): p. 13351-13359.
- [25] Trout, J., *Fraudsters, Churches, Economy, and the Expectations Gap: Applying Trends of Occupational Fraud to an Assurance Engagement Team Plan and Fraud-Prevention Client Proposal*. 2014, University of Mississippi.
- [26] Peiris, M. and U. Rathnasiri, *Detection of Occupational Fraud on Leasing Companies*. 2016.
- [27] Sreejesh, S., S. Mohapatra, and M. Anusree, *Business research methods: An applied orientation*. 2014: Springer.
- [28] Vaccaro, A. and P. Madsen, *Firm information transparency: Ethical questions in the information age*. *Social informatics: An information society for all?* In remembrance of Rob Kling, 2006: p. 145-156.
- [29] Kohlbacher, M. and S. Gruenwald, *Process ownership, process performance measurement and firm performance*. *International Journal of Productivity and Performance Management*, 2011. 60(7): p. 709-720.
- [30] Malinova, M. and J. Mendling. *A qualitative research perspective on BPM adoption and the pitfalls of business process modeling*. in *International Conference on Business Process Management*. 2012. Springer.
- [31] Wells, J.T., *Protect small business*. *Journal of Accountancy*, 2003. 195(3): p. 26.
- [32] Cowan, C., *Software security for open-source systems*. *IEEE Security & Privacy*, 2003. 99(1): p. 38-45.
- [33] Rawlins, B.R., *Measuring the relationship between organizational transparency and employee trust*. 2008.
- [34] Vössing, M., et al., *Designing useful transparency to improve process performance—evidence from an automated production line*. 2019.
- [35] Hultman, J. and B. Axelsson, *Towards a typology of transparency for marketing management research*. *Industrial marketing management*, 2007. 36(5): p. 627-635.
- [36] De Fine Licht, J., *Do we really want to know? The potentially negative effect of transparency in decision making on perceived legitimacy*. *Scandinavian Political Studies*, 2011. 34(3): p. 183-201.
- [37] Huda, S., R. Sarno, and T. Ahmad, *Increasing Accuracy of Process-based Fraud Detection Using a Behavior Model*. *International Journal of Software Engineering and Its Applications*, 2016. 10(5): p. 175-188.

- [38] Reijers, H. and J. Mendling. Modularity in process models: Review and effects. in *International Conference on Business Process Management*. 2008. Springer.
- [39] Laue, R. and A. Gadatsch. Measuring the understandability of business process models-Are we asking the right questions? in *International Conference on Business Process Management*. 2010. Springer.
- [40] Recker, J. and A. Dreiling, Does it matter which process modelling language we teach or use? An experimental study on understanding process modelling languages without formal education. *ACIS 2007 Proceedings*, 2007: p. 45.
- [41] Melcher, J., et al. On measuring the understandability of process models. in *International Conference on Business Process Management*. 2009. Springer.
- [42] Saunders, M.N. and P. Lewis, *Doing research in business & management: An essential guide to planning your project*. 2012: Pearson.
- [43] Design, Q., *How to Plan, Structure and Write Survey Material for Effective Market Research (Market Research in Practice Series)(Paperback)* by Ian Brace; 289 pages. Kogan Page.
- [44] Zugal, S., et al. Assessing the impact of hierarchy on model understandability—a cognitive perspective. in *International conference on model driven engineering languages and systems*. 2011. Springer.
- [45] Dikici, A., O. Turetken, and O. Demirors, Factors influencing the understandability of process models: A systematic literature review. *Information and Software Technology*, 2018. 93: p. 112-129.
- [46] Hair, J.F., et al., *Multivariate data analysis (Vol. 6)*. 2006, Upper Saddle River, NJ: Pearson Prentice Hall.
- [47] Van Der Aalst, W., et al. *Process mining manifesto*. in *International Conference on Business Process Management*. 2011. Springer.

A Tree-profile Shape Ultra Wide Band Antenna for Chipless RFID Tags

A K M Zakir Hossain¹, Nurulhalim Bin Hassim^{2*}, Jamil Abedalrahim Jamil Alsayaydeh³

Centre for Telecommunication Research & Innovation (CeTRI), Fakulti Teknologi Kejuruteraan Elektrik & Elektronik (FTKEE)
Universiti Teknikal Malaysia Melaka (UTeM), Melaka

Mohammad Kamrul Hasan⁴

Center for Cyber Security
Faculty of Information Science and Technology
The National University of Malaysia
Kuala Lumpur, Malaysia

Md. Rafiqul Islam⁵

Department of Electrical and Computer Engineering
Faculty of Engineering
International Islamic University Malaysia
Gombak, Kuala Lumpur, Malaysia

Abstract—In this article, a new small size planar microstrip tree profile shaped Ultra-Wide Band (UWB) antenna with partial ground plane has been presented. The antenna is designed for chipless RFID tags that are working in UWB region. The operating frequency of the antenna is between 2.72 GHz to 11.1 GHz which covers the entire UWB frequency band. The antenna exhibits comparatively high realized gain of 4.2 dBi with respect to its small size of $27 \times 40 \text{ mm}^2$ and have a gain to aperture ratio of 0.243 which is comparatively higher than other existing retransmission-based chipless RFID antennas. Another aspect of this antenna is its total efficiency which never goes below 80% throughout the entire bandwidth whereby it reaches as high as 96% at 3.5GHz. This design will motivate the chipless RFID designers to produce small size and cost effective tags.

Keywords—Planar microstrip; UWB antenna; chipless RFID; realized gain; total efficiency

I. INTRODUCTION

Since the first application of the Radio Frequency Identification (RFID) in WWII to distinguish the friendly airplanes from foe, the RFID has gone through many phases of development and elevated to a level of a dedicated and reliable technology for tagging and Identification (IDing). In application such as livestock management, library management, retail shop, inventory management, toll collection and many others, the RFID system is Omni-present [1]. Based on the availability of on-board power supply on the RFID tags, the RFID can be classified into three classes i) Active (with battery onboard), ii) semi-active/passive RFID (with battery onboard but only to keep the memory chip alive and depends on the reader signal power for communication) and iii) passive RFID (has no battery and fully depends on the reader signal power to perform all functionality). Again, the passive type RFIDs can be sub-divided into two different kinds, a) chipped RFID and b) chipless RFID (CRFID). Among these kinds, the CRFID is inherently cheap due to the absence of the chip [2-3]. The preexisting barcode system is used widely in the industry. However, this system has many problems such as it is vulnerable to wear, tear and dent,

tempering, low security and needs the line of sight (LOS) for detection. Whereas, the CRFID solves all these issues and is the only contender at the forefront to fully replace the barcode system. However, the CRFID still couldn't beat the cheapness of the barcode system. This is due to the comparatively bigger dimensions of the CRFID tag. The CRFID may or may not involve the antennas onboard. However, those CRFID systems that have no antennas on the tag (called backscatter and RCS base CRFID) suffers several issues such as crosstalk problem at the reader and lead to huge difficult to detect [4]. On the other hand, the CRFID with antennas on the tag (called retransmission-based CRFID) solves these issues by using the antenna polarization mismatch technique. It utilizes an antenna with horizontal polarization at the transmitter (Tx) part of the reader which is matched with the receiving (Rx) antenna of the tag. Similarly, the tag's Tx antenna is vertically polarized and matched with the Rx antenna at the reader. This configuration imposes a 90-degree polarization mismatch between the Rx and Tx antennas of both reader and tag; and makes a theoretical zero crosstalk between them. However, the solution comes with a price of making the CRFID tag comparatively bigger than the backscatter/RCS based tag [1]. Fig 1 illustrates the basic working principle of a retransmission based CRFID system.

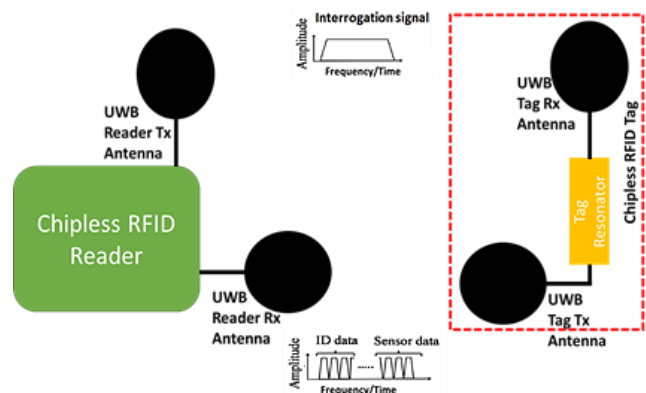


Fig. 1. The Chipless RFID System.

*Corresponding Author

This article is organized as: II. Related work (where the current trends are addressed and benchmarked with this proposal). III. MTPS antenna design (in this section the design procedure is described and elaborated). IV. Results and Discussions (comprises the results and analysis related to the proposed design) and V. Conclusion (where the article is concluded and future recommendation has been made).

II. RELATED WORKS

One of the prominent retransmission-based CRFID has been proposed in [5] which has microstrip disc loaded monopole antennas (DLMA) with a large dimension of $60 \times 66 \text{ mm}^2$ covering the entire bandwidth (BW) of 3.2 GHz to 10.7 GHz (defined UWB by FCC). Despite of the large dimension of that antenna, it only exhibits a low realized gain of 0 dB. Another microstrip DLMA antenna has been proposed in [6] for a CRFID multi-state tag which has a large size of $60 \times 90 \text{ mm}^2$. However, no gain/directivity Vs frequency data has been presented in this work. The same DLMA structure has been proposed in [7], however, using the coplanar waveguide (CPW) feeding for the retransmission based CRFID system on the flexible substrate (Kapton film) to utilize it for liquid concentration sensing. The dimension of the proposed antenna is $40.6 \times 51.1 \text{ mm}^2$, still, with a low gain of 0 dBi. To increase the gain of the antenna, in [4], the DLMA structure has been modified into an elliptical structure, using the same CPW feeding technique. The gain has been observed as 2 dBi, yet, in this particular proposal the actual size of the proposed antenna is missing. In [8], another proposal has been made by using the same CPW feed but the DLMA structure has been modified into a semi-circle structure. The proposed antenna has a comparatively small dimension of $35 \times 32 \text{ mm}^2$. However, the authors have not presented any directivity/gain information in that particular work.

So far, the smallest antenna (a rectangular patch) has been proposed in [9] with a dimension of $23 \times 36 \text{ mm}^2$ and it has a maximum gain of 2.45 dBi for the retransmission-based CRFID tags. However, the gain to aperture (total area used by the antenna) ratio is still low at 0.23. The same rectangular patch UWB antenna has been modified and proposed in [10] for a 3-state CRFID tag. The antenna exhibits a good gain of 4.9 dBi, yet, that is only due to the large dimension ($25 \times 74.62 \text{ mm}^2$) of the antenna by achieving a gain to aperture ratio of only 0.16. Table I summarizes the complete scenario.

TABLE I. SUMMARY OF THE EXISTING WORKS

References	Antenna Size (mm ²)	Gain (dBi)/(Linear)	Gain to Aperture ratio (linear)
[1]	na	na	-
[4]	na	2/1.58	-
[5]	60 × 66	0/1	0.025
[6]	60 × 90	na	na
[7]	40.6 × 51.1	0/1	0.048
[8]	35 × 32	na	na
[9]	23 × 36	2.45/1.75	0.23
[10]	25 × 74.62	4.9/3.01	0.16
Proposed work	27 × 40	4.2/2.63	0.243

*na = not available

For all those aforementioned antennas, some have good gain but due to the large size and other have smaller dimensions, however, suffered a low gain for the chipless RFID system. In this proposed work, a microstrip tree-profile shape (MTPS) antenna has been proposed, designed and simulated for comparatively small size and higher realized gain to overcome these issues.

III. MTPS ANTENNA DESIGN

Fig 2 shows the proposed MTPS antenna structure for the CRFID tag. The antenna has been modeled on the dielectric substrate Rogers RT/ Duroid 5880. This substrate material has a relative permittivity/dielectric constant (ϵ_r) of 2.2, the loss tangent/dissipation factor ($\tan\delta$) of 0.0009 and the dielectric height of 0.508 mm. The antenna has front patch which looks like a tree profile shape and can be seen in the front view in Fig 1. The MTPS patch is connected with a microstrip transmission feed line of 50Ω to the modeled SMA connector. At the back view of the proposed structure, it can be seen that the antenna ground plane has a partial ground plane (PGP) to promote the structure to a UWB antenna geometry.

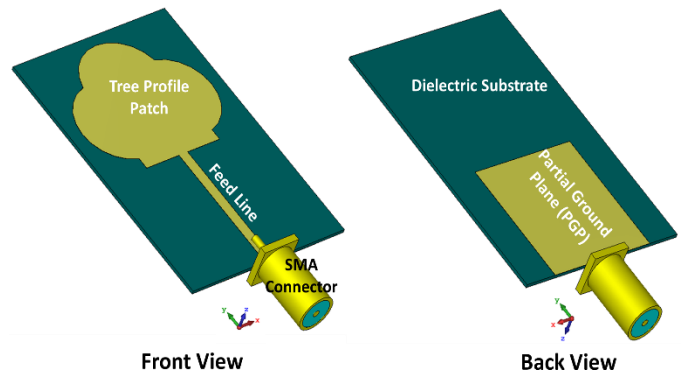


Fig. 2. The Geometry of the Proposed Antenna.

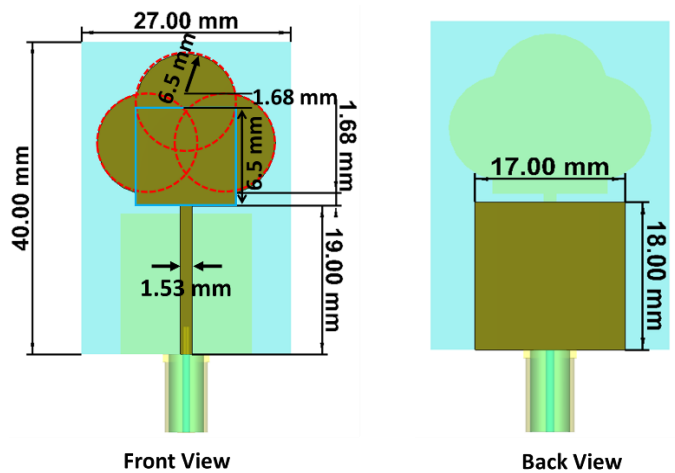


Fig. 3. The Exact Dimensions of the Proposed Antenna.

Fig 3 reveals the exact dimension of the different segments of the proposed MTPS antenna. It can be seen that to create a tree-profile shape it constitutes three differently placed circular patches along with a square patch. All circular patches have the same radius of 6.5 mm whereby the square patch has the side

lengths same as the radius of the circular patch. The radius location of the center circular patch is at the bisection point of the top side of the square but it is 1.68 mm higher from that point towards Y-axis. Similarly, the centers of the other two circular patches are also on the two corresponding sides of the square but 1.68 mm higher from the bisection point of them towards Y-axis. All these four different patches are merged together to form a tree-profile shape. Furthermore, a TL (50 Ω) feed having width of 1.53 mm has been connected with the tree-profile shaped patch that has a length of 19 mm to create a MTPS antenna with a final dimension of 27 × 40 mm². From the back view in Fig 3, it can be seen that the PGP has a length and width of 18 mm and 17 mm respectively.

These dimensions are in fact approximated and later optimized with the help of equation (1) - (9) [11-14].

$$W = \frac{C_0}{2f_c \sqrt{\frac{(\epsilon_r+1)}{2}}} \quad (1)$$

$$\epsilon_e = \frac{\epsilon_r+1}{2} + \frac{\epsilon_r-1}{2} \left[1 + 12 \frac{h}{W_p} \right]^{-1/2} \quad (2)$$

$$L_e = \frac{C_0}{2f_c \sqrt{\epsilon_e}} \quad (3)$$

$$\Delta L = 0.412h \frac{(\epsilon_e+0.3) \left(\frac{W_k}{h} + 0.264 \right)}{(\epsilon_e-0.258) \left(\frac{W_k}{h} + 0.8 \right)} \quad (4)$$

$$L = L_e - 2\Delta L \quad (5)$$

$$L_g = 6h + L_k \quad (6)$$

$$W_g = W_g = 6h + W_k \quad (7)$$

$$r = \frac{F}{\left\{ 1 + \frac{2h}{\pi \epsilon_r F} \left[\ln \left(\frac{\pi F}{2h} \right) + 1.7726 \right] \right\}^{1/2}} \quad (8)$$

$$F = \frac{8.791 \times 10^9}{f_c \sqrt{\epsilon_r}} \quad (9)$$

Where, the L and the W are the length and the width of the square patch for the MTPS antenna respectively. ϵ_e and ϵ_r are the effective and relative dielectric constant respectively; and C_0 is the speed of the light. W_g and L_g are the approximated width and length of the PGP respectively, before optimization. Furthermore, r is the approximated radius of those three circles for the MTPS antenna patch. The approximation has been done by the above stated equations in the beginning. Later, in the CST MWS 2020 simulator, the optimization has been done to achieve the desired BW for the antenna and the final dimensions for the designed MTPS antenna are shown in Fig 3. In addition, the conductors in the design (patch, TL and ground plane) have been modeled as perfect electric conductor (PEC) in the simulator.

IV. RESULTS AND DISCUSSIONS

Fig 4 illustrates three important responses of the designed MTPS antenna; reference impedance, S-parameter and voltage standing wave ratio (VSWR). For any antenna design, the first thing to check is the reference impedance. Since, the antenna will be connected to a system/connector which would have a specific characteristic impedance. In this case, the antenna is

intended to be connected with a 50 Ω system. So, the antenna reference impedance has to be as close as to 50 Ω for a good matching and, less reflection and return loss. Fig 4 (a) shows that the antenna reference impedance is exactly 50 Ω throughout the span of 1 – 12 GHz. Thus, the preliminary design integrity check has been fulfilled.

The next step is to check the S-parameter (S_{11}) response of this antenna to see the preliminary -10 dB BW. From Fig 4 (b) it can be observed that the -10 dB BW for the MTPS antenna starts at 2.75 GHz and ends at 10.95 GHz. So, based on this response a BW of 8.2 GHz is achieved which covers the whole FCC defined UWB frequency. However, to get a confirmation on the actual operating frequency of the antenna, it is necessary to check the VSWR as well. It is known from the theory that a perfect/ideal VSWR for any passive microwave TL is 1. Again, as long as the VSWR value stays between 1-2, the antenna can be operate-able on those frequency band(s). From Fig 4 (c), it can be seen that the antenna has a VSWR of 2.72 GHz to 11.1 GHz. This actually confirms the original working BW of 8.38 GHz for this designed antenna. Fig 5 illustrates the out-band (at 2 GHz) and in-band (at 6.5 GHz) surface current distribution/accumulation of the MTPS antenna. It is seen that at 2 GHz most of the surface current accumulates on the feed line and does not reach on the antenna patch for radiation which implies the antenna cannot radiate good at that frequency. Whereby, at 6.5 GHz it is clearly visible that the current accumulation reaches on the patch of the MTPS antenna for radiation. This is another justification on the quality of the design.

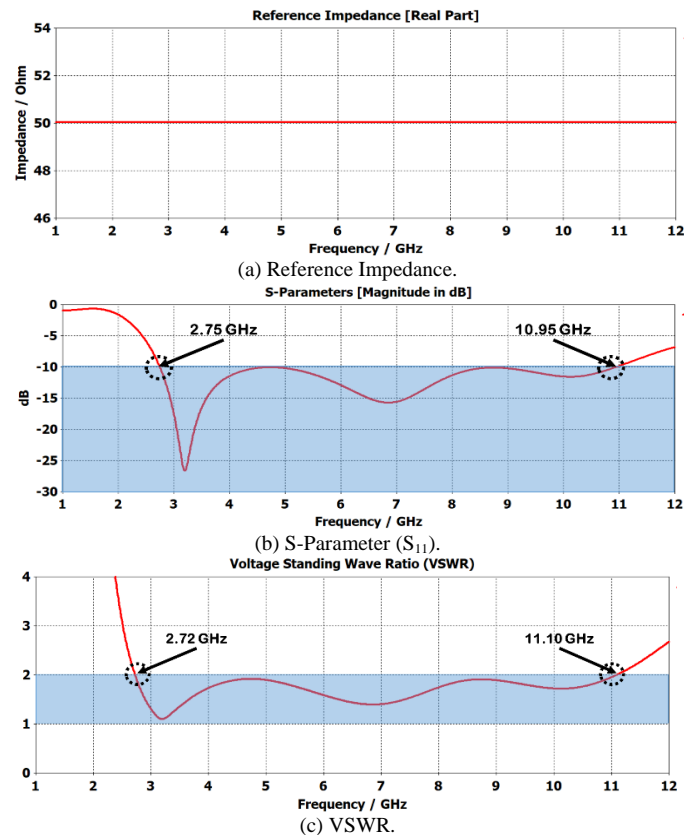


Fig. 4. The (a) Reference Impedance, the (b) S-Parameter (S_{11}) and the (c) VSWR of the Antenna.

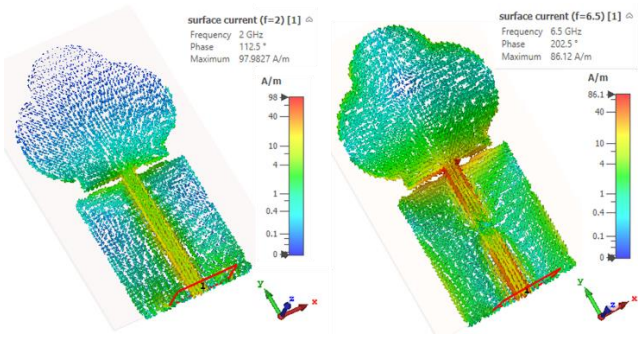


Fig. 5. The Surface Current Distribution at 2 GHz and 6.5 GHz.

Next, it is necessary to investigate the farfield radiation capabilities of the antenna. Fig 6 comprises the 3-D and polar responses of the farfield radiation pattern of the MTPS antenna at 3.2 GHz, 6.5GHz and 10.7 GHz. Fig 6 (a), (c) and (e) illustrate the 3-D patterns for at 3.2 GHz, 6.5GHz and 10.7 GHz respectively. Whereby, Fig 6 (b), (d) and (f) include the polar form of the horizontal (H-) and the elevation (E-) plane pattern at those three frequencies respectively. From the 3.2 GHz response it can be seen that the 3-D pattern is a semi omni-directional pattern where the H-plane is circular and the E-plane is bidirectional in nature. With the increases in the frequency, the farfield pattern becomes more directional as can be seen in Fig 6 (c) and (d). Here, it is observed that the antenna becomes completely bidirectional having pointing lobes at the front and back side of the antenna. The rest of the direction of the antenna, it tends to decrease the intensity of the radiation. From the E- and the H-plane polar representation also justifies the bidirectional nature of the antenna at 6.5 GHz. The response at 10.7 GHz showing that the antenna no longer radiates at the front and back of the antenna structure, rather, it scans the 45° angles from the front and back axis and also radiates prominently at the Z-axis direction. Lastly, it is important also to observe the antenna efficiencies and the realized gain of the antenna. Fig 7 discloses the responses of efficiencies and realized gain vs frequency.

From Fig 7(a), it is realized that both of the efficiency responses are good. The radiation efficiency of the antenna is constant throughout the entire BW between 92% and 99%. Whereby, the total efficiency of the antenna never goes below 80% starting from 2.7 GHz to 11 GHz. Furthermore, it reaches as high as 96% at 3.5 GHz and keeps itself nominally constant around 90% which indicates good quality of the design. Fig 7 (b) comprises the realized gain Vs frequency result of the designed MTPS antenna. It can be seen that the maximum realized gain is 4.2 dBi at 6.7GHz. The gain at the beginning of the bandwidth starts with a value of 2 dBi and reaches at its peak at 6.7GHz and starts falling until 8.8 GHz where the lowest gain of 0.45 dBi can be observed. The gain starts showing rising trends after that point and cross-passes through the working BW with the same trend. All these results presented in this section justify the quality of the proposed MTPS antenna for the retransmission-based CRFID tags and the reader as well.

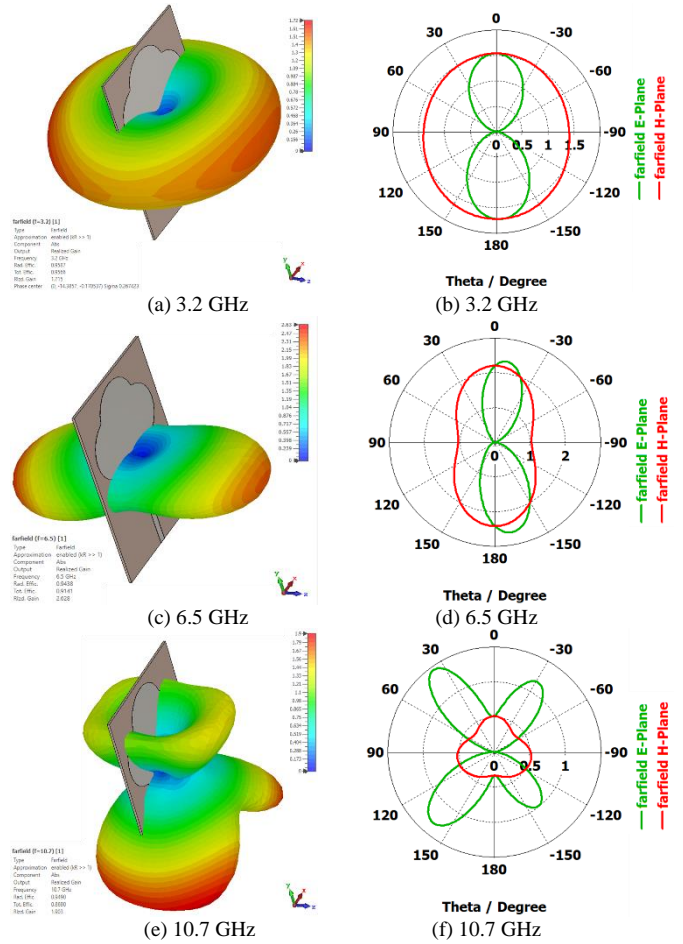


Fig. 6. The 3-D and 2-D (Polar) Representation of the Farfield Radiation Pattern of the MTPS Antenna.

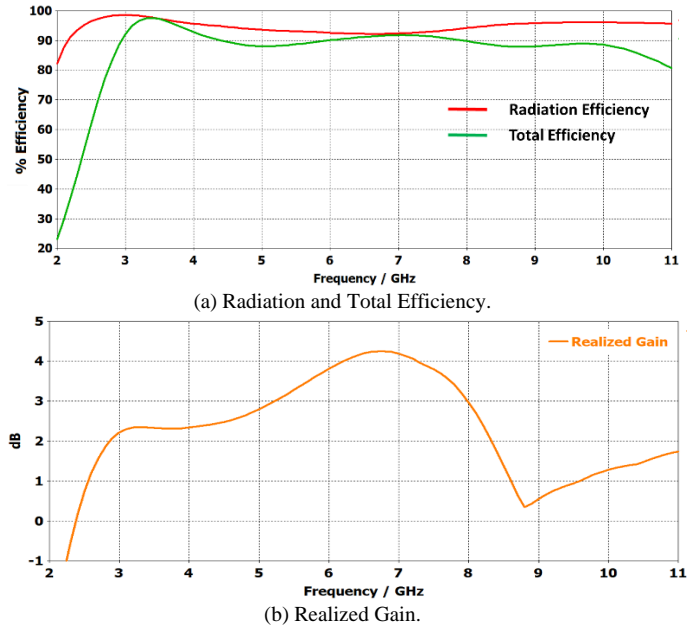


Fig. 7. The (a) Antenna Efficiencies and the (b) Realized Gain Vs Frequency of the Antenna.

V. CONCLUSION

A planar microstrip tree profile shaped (MTPS) UWB antenna for retransmission-based CRFID has been designed and analyzed. The antenna has been modeled in CST MWS and realized on Rogers RT/Duroid 5880. The antenna works between 2.72 GHz to 11.1 GHz which covers the whole UWB frequency allocation. The antenna is small in size, exhibits a good maximum realized gain of 4.2 dBi and has a gain to aperture ratio of 0.243 which is higher than other existing proposed antennas for retransmission-based CRFID applications. The farfield radiation pattern and the efficiencies also indicate the good quality of the design. This proposed antenna will motivate the CRFID researchers to design a comparatively smaller, efficient and cost effective CRFID tags. The future aspect of this work is to implement the proposed antenna with a full CRFID system to assess the capability of this design.

ACKNOWLEDGMENT

This work has been funded by the center for research and innovation management (CRIM), University Teknikal Malaysia Melaka (UTeM).

REFERENCES

- [1] Hossain, A. K. M. Z., Ibrahimy, M. I., Motakabber, S. M. A., Azam, S. M. K., & Islam, M. S. (2021). Multi-resonator application on size reduction for retransmission-based chipless RFID tag. *Electronics Letters*, 57(1), 26-29.
- [2] Hossain, A. K. M. Z., Motakabber, S. M. A., & Ibrahimy, M. I. (2015). Microstrip spiral resonator for the UWB chipless RFID tag. In *Progress in Systems Engineering* (pp. 355-358). Springer, Cham.
- [3] Hossain, A. K., Ibrahimy, M. I., & Motakabber, S. M. (2014). Spiral resonator for ultra wide band chipless RFID tag. In *2014 International Conference on Computer and Communication Engineering* (pp. 281-283). IEEE.
- [4] Bhuiyan, M. S., & Karmakar, N. C. (2014). An efficient coplanar retransmission type chipless RFID tag based on dual-band McSrr. *Progress In Electromagnetics Research*, 54, 133-141. doi:10.2528/PIERC14061403.
- [5] Preradovic, S., Balbin, I., Karmakar, N. C., & Swiegers, G. F. (2009). Multiresonator-based chipless RFID system for low-cost item tracking. *IEEE Transactions on Microwave Theory and Techniques*, 57(5), 1411-1419. doi: 10.1109/TMTT.2009.2017323.
- [6] Majidifar, S., Ahmadi, A., Sadeghi-Fathabadi, O., & Ahmadi, M. (2015). A novel phase coding method in chipless RFID systems. *AEU-International Journal of Electronics and Communications*, 69(7), 974-980. doi:10.1016/j.aeue.2015.02.013.
- [7] Li, Z., & Bhadra, S. (2019). A 3-bit fully inkjet-printed flexible chipless RFID for wireless concentration measurements of liquid solutions. *Sensors and Actuators A: Physical*, 299, 111581. doi:10.1016/j.sna.2019.111581.
- [8] Weng, Y. F., Cheung, S. W., Yuk, T. I., & Liu, L. (2013). Design of chipless UWB RFID system using a CPW multi-resonator. *IEEE Antennas and Propagation Magazine*, 55(1), 13-31. doi: 10.1109/MAP.2013.6474480.
- [9] Ma, Z. H., Yang, J. H., Chen, C. C., & Yang, C. F. (2018). A retransmitted chipless tag using CSRR coupled structure. *Microsystem Technologies*, 24(10), 4373-4382. doi:10.1007/s00542-018-3836-z.
- [10] Abdulkawi, W. M., & Sheta, A. A. (2020). High coding capacity chipless radiofrequency identification tags. *Microwave and Optical Technology Letters*, 62(2), 592-599. doi:10.1002/mop.32057.
- [11] Islam, M. S., Ibrahimy, M. I., Motakabber, S. M. A., & Hossain, A. Z. (2018, September). A Rectangular Inset-Fed Patch Antenna with Defected Ground Structure for ISM Band. In *2018 7th International Conference on Computer and Communication Engineering (ICCCE)* (pp. 104-108). IEEE. doi:10.1109/ICCCE.2018.8539260.
- [12] Islam, M. S., Ibrahimy, M. I., Motakabber, S. M. A., Hossain, A. Z., & Azam, S. K. (2019). Microstrip patch antenna with defected ground structure for biomedical application. *Bulletin of Electrical Engineering and Informatics*, 8(2), 586-595. doi:10.11591/eei.v8i2.1495.
- [13] Azam, S. M. K., Islam, M. S., Hossain, A. K. M. Z., & Othman, M. (2020). Monopole antenna on transparent substrate and rectifier for energy harvesting applications in 5G. *International Journal of Advanced Computer Science and Applications*, 11(8), 84-89. doi:10.14569/IJACSA.2020.0110812.
- [14] Zakir Hossain, A. K. M., Hassim, N. B., Kayser Azam, S. M., Islam, M. S., & Hasan, M. K. (2020). A planar antenna on flexible substrate for future 5g energy harvesting in malaysia. *International Journal of Advanced Computer Science and Applications*, 11(10), 151-155. doi:10.14569/IJACSA.2020.0111020.

Non-Hodgkin Type Lymphoma Cancer Cell Detection using Connected Components Labeling and Moments of Image

Monirul Islam Pavel¹, Mohsinul Bari Shakir², Dewan Ahmed Muhtasim³, Omar Faruk⁴

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology^{1,3}

The National University of Malaysia, Bangi, Selangor, Malaysia

Department of Computer Science and Engineering, BRAC University, Dhaka, Bangladesh²

Faculty of Information Science and Technology, The National University of Malaysia, Bangi, Selangor, Malaysia⁴

Abstract—Cancers are one of the deadliest diseases with a costly treatment system in the world at present. In this paper a cost-effective, autonomous system of cancer-cell detection was proposed using several efficient image processing methods to develop an early stage non-Hodgkin type lymphoma which is a type of blood cancer. The system is implemented automatically to detect the traits of cancer in microscopy images of biopsy samples. Recent attempts have previously lacked flexibility in characteristics and the accuracy level is not consistent with the individual cancer type. The framework consisted of three stages for detecting cancer on the basis of various detected traits including cell segmentation, quantification, area measurement analysis of cells, a center clump detection using the moment of image, identification of 4-connected components and Moore-Neighbor tracing algorithm. This methodology has been used in several sets of images and Feedback from these test executions has been used to improve the system. Subsequently, the proposed method can be used efficiently for used for autonomous non-hodgking type lymphoma cancer cell detection, which has an accuracy of 93.75%.

Keywords—Non-hodgking; lymphoma; moment of image; connected components labeling; Otsu thresholding

I. INTRODUCTION

The term cancer is referred to as a barrier to anomalous cell division. Cancer cells can migrate across blood, lymph systems and tumors to other areas of the body. But all tumors are not cancerous, tumors may be benign as not cancerous, or malignant (cancerous). There are over one hundred types of cancer that has been recognized and each type has many subtypes which has variations of their own. This immense variation makes cancer detection very complex, especially in the preliminary stages. The causes of cancer, in most cases, are still not very well understood. Hence treating cancer becomes even more strenuous [1-3]. Due to this enormous complexity of the disease scientists, doctors and engineers all over the world are researching on the field of cancer to achieve a better understanding of cancer and find absolute cures for each type of cancer in the process. Even though the process is lengthy and difficult, but knowing more will enable doctors to cure cancer patients more effectively. This motivated us to think about the mechanism of cancer detection and use technology to speed up the process. If cancer researchers are able to automatically detect cancer cells via means of image

processing, this can save tremendous amounts of time and also increase efficiency of the research, since the human error factor will cease completely. For detail clarification of implementation techniques, several similar works had been studied throughout the research. For instance, an automatic detection method was introduced by Agaian, S. et al. for Acute Myelogenous Leukemia where 80 microscopic image data, collected from the American Haematology Society, were used. The authors used k-mean cluster algorithms to extract the nuclei of the cell in the pre - processing phase. Then extraction of features by Hausdorff Dimension was carried out to count the number of the boxes. SVM is then adapted as the classification where the accuracy is 98% [4]. Two approaches to classify blood cell cancer were suggested based on the doctor's guide by separating L1, L2, M5 AML and comparing with other forms of leukemia. The working architecture was developed based on the Gaussian distribution and Random Forest Classification methodology, after transforming RGB to YCbCr color space and this solution was able to get 94 percent accuracy [5]. Leukemia detection with leucocytes classification was performed by Putzu, L. et al. [6] using image processing techniques including color conversation, contrast stretching, applying Zack Algorithm for segmentation, removing backgrounds and so on. Total seven types of feature extraction calculation were applied including measurement of roundness, convexity, compactness, elongations, eccentricity, rectangularity, and these were fed SVM classifier where the accuracy was more than 80% deploying on 33 test images.

II. THEORITICAL STUDIES

A. Connected Components Labeling (CCL)

The CCL scans an image depending on pixel connections to identify connected areas. Connect component pixels are somehow related where they are bound to those pixel intensity values. Both pixels are marked with a color after evaluating the connected areas. Connected labeling components searches an image from top to bottom and from left to right, pixel-by-pixel, for instance, to distinguish areas of neighbor pixels of same intensity values [7].The CCL operator scans the image by moving along till it complete a loop from the coordinates, it found and started. It thoroughly scans like its own process with the concept of four connections. If any non-zero pixel (white) is found, it starts the loop and stores the pixels that at least have

one connection with its neighbor one. Through this process, a certain pixels of bounded area are stored. To obtain the contour pixels, while the scanning in an image, if any of four connected components is missed, it stores as a contour pixels.

B. Douglas-Peucker Algorithm

Douglas-Peucker is necessary to have the line segments approximate the initial direction. In topology, the ultimate simpler path is compatible with the initial path, in particular with neighborhood trajectory properties. The characteristic points are extracted and the original trajectory, approximating the original trajectory, is then reconstructed. The benefit of the fundamental DP is that the measuring outcome is definite when the curve and threshold are specified, with a rotation and translation entropy. In order to optimize the rows, the threshold must be predetermined by the users [8]. All points are illustrated from the first to the last stage as well as the first and the last stage are automatically retained. The point is the one with first and last points as nodes that are the further off from the section with the curve nodes, where one point is similar to the line section than epsilon, all items that have not actually been defined are removed without an aided scale of the worst than epsilon. When the furthest point is bigger than an epsilon from the line segment, the point is kept. The approach applies frequently to the first, and furthest and then the last, which includes the distance marked as conserved. A new curve with the values labeled for retention after completion of the incident is created.

C. Moore's Neighbor Tracing Algorithm

The pattern group of white pixels in this Algorithm is positioned on a black pixel backdrop. It is taken as the starting pixel when a white pixel came at the left end of the pixel range. Afterwards, the contour was extracted from this pixel in a clockwise direction by moving round the pattern. This enabled machine to map the entire pixel array. The key thought is to go back until the last white pixel backtracking from it hits a black pixel on every time. When the second visit was made to first pixel, the algorithm stops.

III. PROPOSED METHODOLOGY

The following methods are implemented in order to count normal and cancerous cells, segmenting cancerous cells by calculating areas and measure centers and distances between each of those kinds of cells that forms clumps. Fig. 1 described the work flow.

A. Pre-processing

In this proposed model, Non-Hodgkin Lymphoma typed cancer cells' biopsy image in Fig. 2.a that has 1000 x 741 resolutions, is used for analysis [9]. The sample image undergoes with gray scaling shown in Fig. 2.b, to minimize color complexity, converted to binary image using Otsu [10-12] threshold with normalized intensity value 0.55, shown in Fig. 2.c and then image is processed into inverted binary image to make easier while calculating moments of image in the next step to measure area which is visualized in Figure 2.d. Followed by median filtering to reduce noises in Fig. 2.e and flood fill operation [13,14] is applied to fill the background regions using morphological reconstruction which recover the

minima shown in Fig. 2.f that are not connected with in an object boundary.

B. Image Segmentation

As in some cases of cancer, cells don't get bigger or form clump but increase in number by dividing, then the following cell counted method will be applied and also implemented to count and extract only cancerous cells based on size. First of all, boundaries of each cells are traced based on the connectivity's of white pixels from black pixels background. Moore-Neighbor tracing algorithm's modified version Jacob's stopping criteria [14], [15], [16] is applied that scans starting from left bottom left corner to each rows going upwards and again starting from leftmost column to right until stop from where it started If it can complete a loop, then it'll be traced as segmented boundary and quantified as cell .In the end, the final image in binary and RGB form will be displayed by plotted each outer shell of cells marked with green color.

C. Cells' Area Measurement Technique

Cancerous cells enlarge in size in some types including lymphoma. The paper presents two different ways to calculate area of cells, one is based on region extractions and another is by calculating the moments of image.

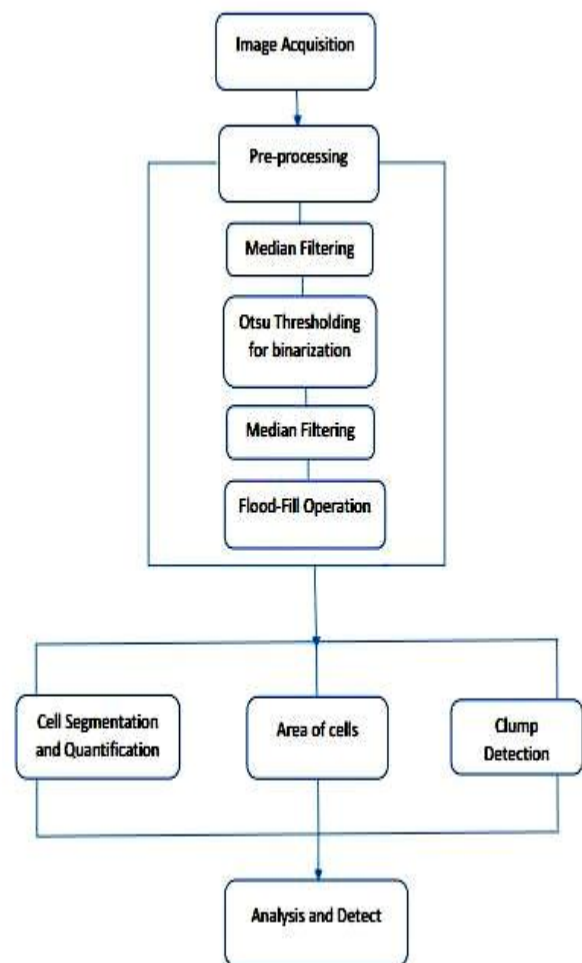


Fig. 1. Proposed Workflow.

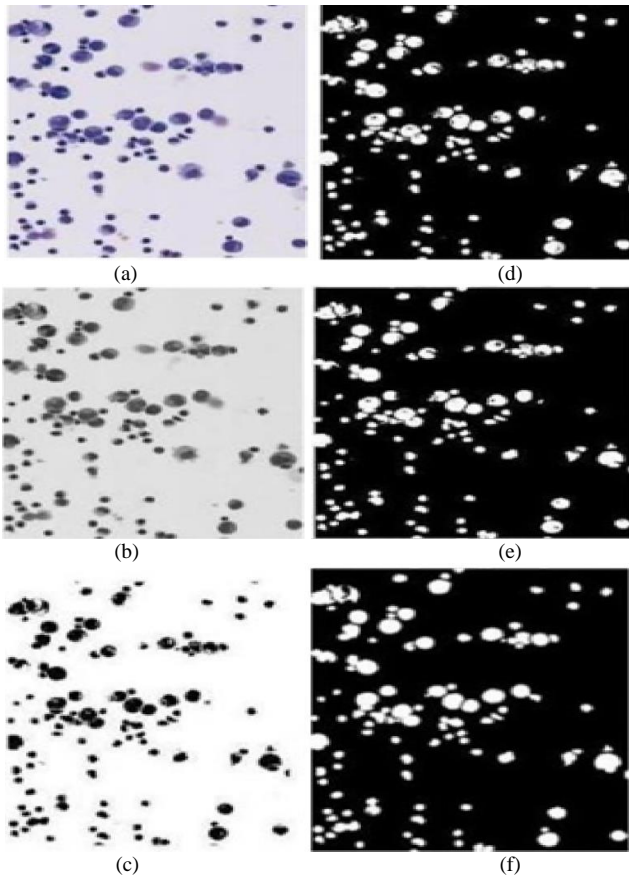


Fig. 2. (a) Sample Image, (b) Gray Scaled, (c) Binarizing using Otsu Threshold 0.55, (d) Inverse Binarize (e) Median Filtered, (f) Flood-Fill Operation.

The method of connecting components are used to find connected regions [17, 18] and how many pixels are connected. As it is 2D image and need to find pixels that are directly connected or touches edges, 4-connected neighborhood are used. Now, in order to measure properties of the regions of connected non-zero pixels, the function of MATLAB named regionprops [19, 20] is implemented. This Returns a quaternion that indicates the actual pixel number in the area. The neighbour regions are the cells. In the connected area the number of non-zero white pixels is the cell region.

In the second proposed method of cells area measuring, regions are extracted and computed after calculating moments of image based on contour approximation method as cells are in irregular shapes. Contour approximate method [21] which is the implementation of Douglas-Peucker Algorithm [22], stores all contour vector points of horizontal, vertical, and diagonal segments using OpenCV's chain_approx_none [23]. If the image is considered as $f(x,y)$ and i, j are any number to calculate image with pixel intensities then the moment of image M_{ij} can be calculated using following equation:

$$M_{ij} = \sum_x \sum_y x^i y^j f(x,y) \quad (1)$$

In the binary image of cells, the zeroth moment M_{00} is the Area [24].

$$A = M_{00} \quad (2)$$

As the binary image's pixels are 1 and 0, if x and y are 0 that means for every white pixels, a '1' will be summed. This process will continue until returning to the starting point of scanning in a connected region. When w and h denote width and height of the image, equation of Area A can be written as below where x^0, y^0 is removed as it doesn't affect the equation.

$$A = \sum_{x=0}^w \sum_{y=0}^h x^0 y^0 f(x,y) \quad (3)$$

As the area of objects is relay on pixel, so the area of same sample may vary based on image resolution. To avoid errors for pixel pitches following equation of ratio is implemented in both area measuring method, while coordinates $x = 1000$ and $y = 741$.

$$Ratio = \frac{New\ Resolution}{x*y} \quad (4)$$

$$A = A * Ratio \quad (5)$$

D. Clump Detection based on Center and Distance Measuring

In some cases of cancer, cells don't enlarge or divided to increase in number, it forms clump. To detect clump, system needs to detect either the affected nucleus or center of each cells. In this sample type of cancer cells image, clump will not occur but method has been applied on to detect center and distance from each. To detect center step of area measuring using moments of image based on contour approximation [25] is followed to get the area using Otsu binarization thresholding (Fig. 3.a) and canny edge detection (Fig. 3.b). Then divide each by the number of pixels that is the zeroth moment ($M_{0,0}$) which is the area of the particular bounded region. Considering sum_x and sum_y are the total x and y coordinates of white pixels, and M_{10}, M_{01} are the calculated average position of each axes.

$$M_{10} = \frac{sum_x}{M_{00}} \quad (6)$$

$$M_{01} = \frac{sum_y}{M_{00}} \quad (7)$$

$$Center = \left(\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right) \quad (8)$$

Here, coordinates of center x and y are described as the spatial moments of first order and dividing with area. the location of centers is marked with RGB value (24,16,247). After that, the distance from one center point to other is measured using distance formula.

$$Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (9)$$

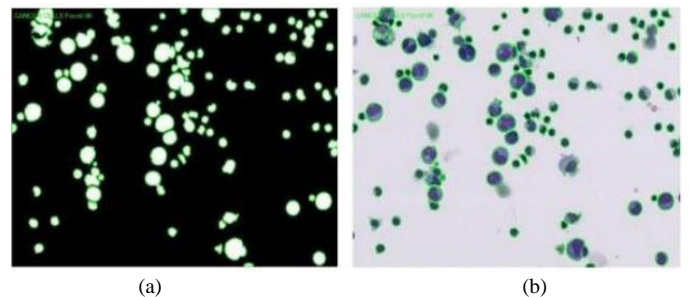


Fig. 3. Counted Cells Plotted in (a) Binary Image, (b) RGB Image.

To reduce errors, lowest three distances are taken, bubble sort algorithm is applied. Afterwards, it is proposed to mark the center of the cells and measure the positions and find out the lowest three distances from one another to reduce error, based on the values of the distances. Hence find out if a clump of cell is present. This system of center measuring is also based on the calculation of the moments of the image [26, 27, 28]. Gaussian smoothing is applied just before the median filtering during pre-processing. Both Otsu algorithm and canny edge detection [29, 30] algorithm are applied to analysis severally and outcomes are visualized in Fig. 4. To split the touching or overlapping cells, watershed algorithm is used based ride line after getting value from distance transformation. Then the cells are contoured. Contour contains the coordinates of cells that have the same outline intensity with fewer number of vertices. Contour approximate method is the implementation of Douglas-Peucke Algorithm [31], it stores all contour vector points of horizontal, vertical, and diagonal segments with opencv'schain_approx_none method. Moments are determined using Discrete Green's Theorem which are the specific weighted average intensities of the image pixels and the boundaries can be drawn using draw contour function of OpenCV.

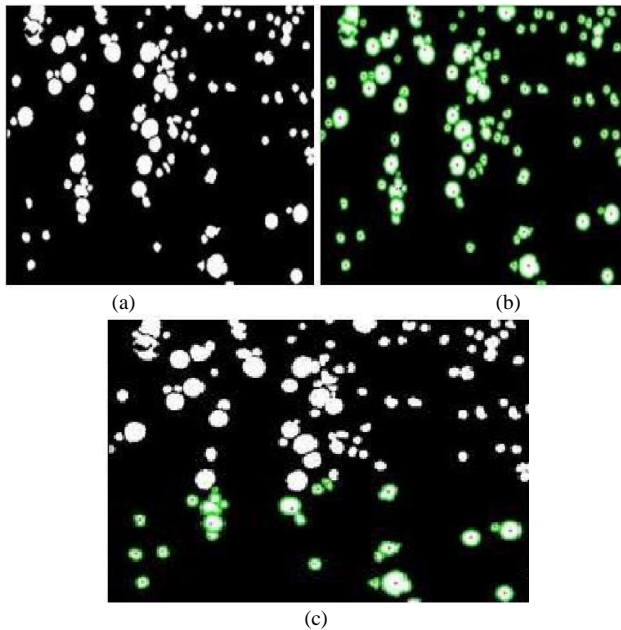


Fig. 4. (a) Pre-processed Image, (b) Center Marked of each Cell using Otsu Threshold, (c) Center Marked using Canny Edge Detection.

IV. RESULT AND DISCUSSION

By doing a relevant study, it is acknowledged that non-Hodgkin type lymphoma cells grows in numbers and enlarge in size. As over lapping is needed to be reduces as much as possible, the system varies on Otsu threshold with normalized intensity value to detect perfect shapes. TABLE I show that while the threshold value is 0.55, all cells are detected in best shapes and total 96 cells are counted throw this system where the actual number of cells is 94 counted manually. So, it gives 97.91% accuracy. Finally based on the threshold value of area gained by trial and error method, total 31 cancerous cells which

is shown in TABLE II, are detected and counted in the particular region of the sample image.

The cells are measured and marked using connected components analysis with a threshold value of 1100 pixels. The threshold value was fine-tuned for this particular image by trial and error method. This process displays out each cell individually with the respective area.

After analysis the cells the following graph Fig. 5 is generated and plotted based on areas. The line above the red line shows the cancer cells and cells plotted below the red line indicates the normal cells. The detected and segmented each cancer cells are shown in Fig. 6. and visualization of their calculated area are displayed in Fig. 7.

Implementation of Otsu binarization threshold produced a more accurate result during center marking for detecting clumps compared to using Canny Edge Detection. TABLE III displayed the cell marking and accuracy difference of both methods.

TABLE I. TOTAL CELLS BASED ON THRESHOLD VALUE

Otsu Threshold value (Binarization)	Total cell	Accuracy (%)
0.1	0	0
0.2	39	41.49
0.2	67	71.27
0.3	105	89.52
0.35	109	86.23
0.4	101	93.06
0.45	101	93.06
0.5	98	95.91
0.55*	96	97.91
0.6	90	95.74
0.65	86	91.48
0.7	80	85.1
0.8	71	75.53
0.85	1	1.06

TABLE II. TOTAL CELLS BASED ON THRESHOLD VALUE

Area (in Pixel)	Cancer Cell Count
700	35
800	33
900	33
1000	33
1050	33
1100	31
1150	29
1200	28
1300	28
1400	24
1500	24

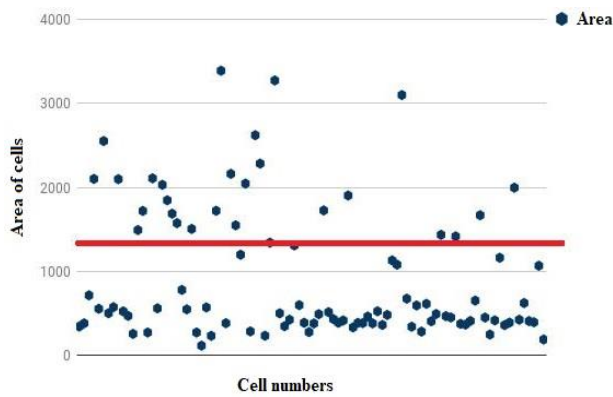


Fig. 5. Area of Cells vs. Segmented Cells.

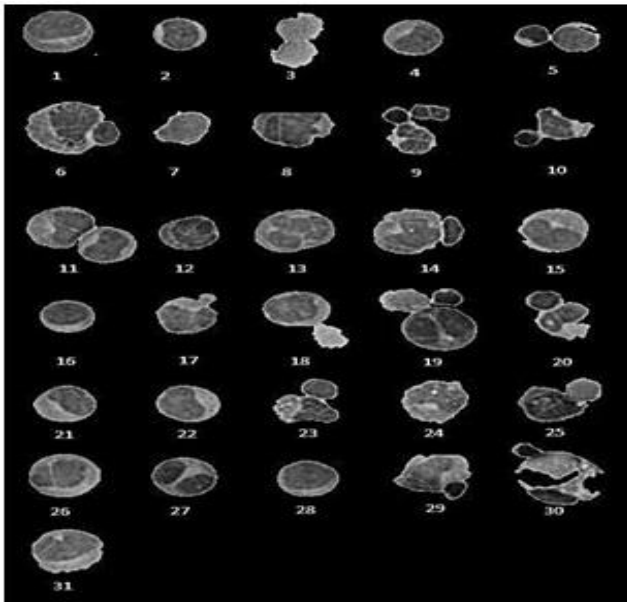


Fig. 6. Detected and Segmented Cancer Cells.

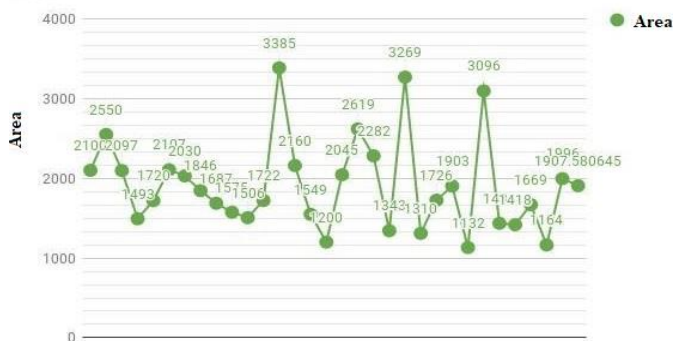


Fig. 7. Visualization of Area of Cancer Cells.

The proposed method is compared with existing three other methods which is shown in TABLE IV. Among the three other methods, algorithms like k-mean cluster, support vector machine (SVM), Zack algorithm for thresholding, watershed segmentation, Fuzzy c-means as well as feature extractors like Gabor Gray Level Co-Occurrence Matrix (GLMC).

TABLE III. ACCURACY TEST OF ALGORITHMS TO FIND CENTERS FOR CLUMP

Threshold using in Center Marking	Cells marked (out of 94)	Accuracy (%)
Otsu Binarization Thresholding	90	93.75
Canny Edge	22	22.91

TABLE IV. COMPARISON WITH EXISTING WORKS

Reference	Method	Accuracy (%)
[32]	Fuzzy c-means clustering, Gabor texture extraction and SVM	90%
[33]	K-mean clustering, GLCM Features	83.33%
[34]	Threshold with Zack Algorithm with Watershed Segmentation and SVM	93%
Proposed Method*	CCL and Moment of Image with Otsu Thresholding	93.75%

V. CONCLUSION

Every year millions of people all over the world are being suffered from cancer and a large percentage of these people die because there is no solid cure to the type of cancer that affected them. Numerous scientific communities are constantly researching on different grounds of cancer to figure out possible defined cures. The proposed model effectively identifies cancer cells simultaneously by cell counts, cell area measurements and clump detection in this study. In terms of traits and the accuracy of 93.75%, the proposed non-hodgking cancer cell detection technology method from microscopic biopsy images may be an optimum approach.

REFERENCES

- [1] Fodde, R. and Brabletz, T., 2007. Wnt/ β -catenin signaling in cancer stemness and malignant behavior. *Current Opinion in Cell Biology*, 19(2), pp.150-158.
- [2] Liu, Y. and Barta, S., 2019. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment. *American Journal of Hematology*, 94(5), pp.604-616.
- [3] The Development and Causes of Cancer. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK9963> (accessed April 15, 2017)
- [4] Agaian, S., Madhukar, M., & Chronopoulos, A. T. (2014). Automated screening system for acute myelogenous leukemia detection in blood microscopic images. *IEEE Systems journal*, 8(3), 995-1004.
- [5] Mohamed, H., Omar, R., Saeed, N., Essam, A., Ayman, N., Mohiy, T., & AbdelRaouf, A. (2018, March). Automated detection of white blood cells cancer diseases. In *Deep and Representation Learning (IWDRDL), 2018 First International Workshop on* (pp. 48-54). IEEE
- [6] Putzu, L., Caocci, G., & Di Ruberto, C. (2014). Leucocyte classification for leukaemia detection using image processing techniques. *Artificial intelligence in medicine*, 62(3), 179-191.
- [7] Mercy, S. S. G., Muthulakshmi, I., & Scholar, P. G. (2018). Automatic number plate recognition using connected component analysis algorithm. *International Journal For Technological Research In Engineering*, 5(7).
- [8] Liu, J., Li, H., Yang, Z., Wu, K., Liu, Y., & Liu, R. W. (2019). Adaptive Douglas-Peucker algorithm with automatic thresholding for AIS-based vessel trajectory compression. *IEEE Access*, 7, 150677-150692.
- [9] Available online : <http://lymphomapictures.org/p/37/non-hodgkin-lymphoma/picture-37> (accessed April 29, 2017)
- [10] Alam, M. A., Shakir, M. B., & Pavel, M. I. (2019, May). Early detection of coronary artery blockage using image processing: segmentation,

- quantification, identification of degree of blockage and risk factors of heart attack. In *Micro-and Nanotechnology Sensors, Systems, and Applications XI* (Vol. 10982, p. 109820L). International Society for Optics and Photonics.
- [11] Zhang J., Hu J, Image Segmentation Based on 2D Otsu Method with Histogram Analysis, *Computer Science and Software Engineering*, 2008 International Conference on Hubei, China, 2008.
- [12] Alam, M. A., Shakir, M. B., Hossain, M. A., Pavel, M. I., Shams, K. M. A., & Akib, F. R. (2018, March). Early detection, segmentation and quantification of coronary artery blockage using efficient image processing technique. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications* (Vol. 10579, p. 105791J). International Society for Optics and Photonics.
- [13] Soille, P., *Morphological Image Analysis: Principles and Applications*, Springer-Verlag, pp. 173-174, 1993.
- [14] Rumman, M., Tasneem, A. N., Farzana, S., Pavel, M. I., & Alam, M. A. (2018). Early detection of Parkinson's disease using image processing and artificial neural network. In *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 256-261).
- [15] Biswas, S., & Hazra, R. (2018). Robust edge detection based on Modified Moore-Neighbor. *Optik*, 168, 931-943.
- [16] Mandeel, T. H., Ahmad, M. I., Isa, M. N. M., Anwar, S. A., & Ngadiran, R. (2018). Palmprint Region of Interest Cropping Based on Moore-Neighbor Tracing Algorithm. *Sensing and Imaging*, 19(1), 15.
- [17] Krishnan G.S.S., Vijaya N., Algorithm on tracing the boundary of medical images using abstract cellular complex, *Machine Vision and Image Processing (MVIP)*, International Conference on Taipei, Taiwan, 2013.
- [18] Chai B., VassJ., Zhuang X., Significance-linked connected component analysis for wavelet image coding " *IEEE Transactions on Image Processing*, Volume: 8, Issue: 6, 1999.
- [19] Zhang, L., & Yu, W. (2017). Orientation image analysis of electrospun submicro-fibers based on Hough transform and Regionprops function. *Textile Research Journal*, 87(18), 2263-2274.
- [20] Ananthanarasimhan, J., Leelesh, P., Anand, M. S., & Lakshminarayana, R. (2020). Validation of projected length of the rotating gliding arc plasma using 'regionprops' function. *Plasma Research Express*, 2(3), 035008.
- [21] Abdelsamea, M. M., Gnecco, G., & Gaber, M. M. (2015). An efficient Self-Organizing Active Contour model for image segmentation. *Neurocomputing*, 149, 820-835.
- [22] Rakun E., AndrianiM., Wiprayoga I.W., Combining depth image and skeleton data from Kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia]), *Advanced Computer Science and Information Systems (ICACISIS)*, 2013 International Conference on Bali, Indonesia, 2014.
- [23] Suzuki, S. and Abe, K., *Topological Structural Analysis of Digitized Binary Images by Border Following*. *CVGIP* 30 1, pp 32-46, 1985.
- [24] Pavel, M. I., Sadique, A. M., Ritul, R. A., Khan, S., & Nath, S. (2017). Cancer detection using image processing techniques based on cell counting, cell area measurement and clump detection, B.Sc. Thesis, BRAC University.
- [25] Y. Yuan, G. Cheung, P. Frossard, P. Le Callet and V. H. Zhao, Contour approximation & depth image coding for virtual view synthesis, *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, Xiamen, pp. 1-6, 2015.
- [26] Python and Opencv: Finding Opencv Contours with Cv2.findcontours. Available online <https://www.pyimagesearch.com/2014/04/21/building-pokedex-python-finding-game-boy-screen-step-4-6/> (accessed July 11, 2017)
- [27] Image Moments . Available online: <http://aishack.in/tutorials/image-moments/> (accessed June 02, 2017)
- [28] Yao, G., Hunte, K., & Dani, A. (2018, June). Image moment-based object tracking and shape estimation for complex motions. In *2018 Annual American Control Conference (ACC)* (pp. 5819-5824).
- [29] Rong, W., Li, Z., Zhang, W., & Sun, L. (2014, August). An improved CANNY edge detection algorithm. In *2014 IEEE International Conference on Mechatronics and Automation* (pp. 577-582).
- [30] Parthasarathy, G., Ramanathan, L., Anitha, K., & Justindhas, Y. (2019). Predicting Source and Age of Brain Tumor Using Canny Edge Detection Algorithm and Threshold Technique. *Asian Pacific journal of cancer prevention: APJCP*, 20(5), 1409.
- [31] Yu, J., Chen, G., Zhang, X., Chen, W., & Pu, Y. (2013, June). An improved Douglas-Peucker algorithm aimed at simplifying natural shoreline into direction-line. In *2013 21st International Conference on Geoinformatics* (pp. 1-5).
- [32] Karthikeyan, T., & Poornima, N. (2017). Microscopic image segmentation using fuzzy c means for leukemia diagnosis. *Leukemia*, 4(1), 3136-3142.
- [33] Gajul, Y. A., & Shelke, R. (2016). Computerized Detection System for Acute Myelogenous Leukemia in Blood Microscopic Images. *Int. J. Innov. Res. Sci. Eng. Technol.*
- [34] Putzu, L., Caocci, G., & Di Ruberto, C. (2014). Leucocyte classification for leukaemia detection using image processing techniques. *Artificial intelligence in medicine*, 62(3), 179-191.

Grey Clustering Method for Water Quality Assessment to Determine the Impact of Mining Company, Peru

Alexi Delgado¹, Jhoel Andy Gauna Achata², Jorge Alfredo Barreda Valdivia³
Julio Cesar Junior Santivañez Montes⁴, Chiara Carbajal⁵

Mining Engineering Section, Pontificia Universidad Católica del Perú, Lima, Peru^{1, 2, 3, 4}
Administration Program, Universidad de Ciencias y Humanidades, Lima, Perú⁵

Abstract—Mining operations have a significant impact on environment, where the quality of water is an important affected issue that needs to be controlled. In that way, the Grey Clustering Method based on center-point triangular whitenization weight (CTWF), is an artificial intelligence criterion that evaluates water samples according to selected parameters, in order to realize an effective water quality assessment. In the present study, the analysis is made on the Crisnejas River Basin, by using fifteen monitoring points based on an investigation realized by the National Water Authority (ANA) in 2019, based on the Peruvian law (ECA) about water quality standards. The results reveal that almost all of the monitoring points on the Crisnejas River Basin were classified as “irrigation of vegetables unrestricted”, but only one point was classified as “animal drink”, which is located in an urbanized area. This implies that mining discharges are being well treated by the company, but another deal is the contamination generated in towns. Further, the present study might be helpful to audit processes made by the state or companies, to justify the quality of surface waters using a more accurate methodology.

Keywords—Grey clustering method; mining company; water quality; artificial intelligence

I. INTRODUCTION

Yanacocha is a well-known mining district in Cajamarca, Perú, characterized by an extensive disseminated gold mineralization [1]. Therefore, the operations and processes used to obtain this metal have an important environmental impact over superficial waters like rivers, which are an important source for people, livestock and plantations [2]. Consequently, there is a necessity to analyze the water quality to assure that concentrations of diverse contaminants are in the range of permitted limits.

For this reason, the need of using a grey clustering evaluation method is proposed, as used in other superficial water quality analysis for rivers [3] [4]. The method is based on grey systems theory, which overcomes the problem about the lack of information of evaluated objects [5]. Also, there is an importance in determining the whitenization weight function; and in that way, the center-point triangular weight function (CTWF) is applied [6].

Yanacocha is located in Cajamarca at 3600 m.a.s.l., in the northern part of Perú. In addition, is at the top of the Crisnejas River Basin [7], which delivers a watershed system

downstream the rivers that cross the city of Cajamarca and communities that live nearby dedicated to livestock [8]. So, the transportation agents are rivers, that depending of their shape, flow rate, width, drainage area and topography, reduce or concentrate a varied record of contaminants.

For the study, we used 15 monitoring points taken from field by ANA [7], and Peruvian law [9]. In that way, the specific objective was to redefine a more accurate classification based on grey clustering, applied to the Crisnejas River Basin, were the results obtained using this methodology determine the grade of negative impact of upstream mining. Furthermore, this study can help people that live along the watershed to have more control over water management.

The structure used for the study is developed as follow: literature review, methodology of the grey clustering system, case of study, results – discussion and conclusions.

II. LITERATURE REVIEW

Grey Systems focuses on the effective processing of available data, in order to deal with the uncertainty, they present. In this way it is necessary to lay a firm methodological foundation for the scientific paper since water quality monitoring and assessment is highly influenced by uncertainty [10]. Other paper about the quality of the Rio Cau provided a way to characterize a river in order to analyze the decision-making process. Furthermore, the Crisnejas River Basin assessment can serve as a model for other mining projects in terms of water resource management [11].

In “The Use of Grey Systems Theory to Analyze the Water Supply Systems Safety”, the use of the grey clustering method based on the grey systems theory is made to evaluate water quality with artificial intelligence criteria. This is due to the fact that in other papers, the limitations of data collection for water supply companies in order to extend the analysis of the matrix, risk in water safety plans [12]. The evaluation uses the monitoring offered published by the National Water Authority (ANA) and the parameters established by MINAM-Peru (DS N° 015-2015), which will be taken into account for this work. The study will focus on the evaluation of the quality of the Santa River and how it can be used for the consumption of the population through different types of water treatment. On the contrary, our study will focus on the management of the water resources of the Yanacocha mine based on the evaluation of

the water quality of the Crisnejas River Basin with the purpose of being a model for the current and future mining projects nearby [5].

“Water Quality Assessment using the Grey Clustering Analysis on a river of Taxco, Mexico” is a research which evaluates the impact of wastewater from a mine on a river in Taxco (Mexico) and how it has been impacted by the mining activity. For this reason, the grey clustering classification method was used to evaluate water samples in 4 different points. Also, this work shows how the proximity to the mining operation impact in the degree of contamination. Another study which is similar as the latter, is “Water Quality in Areas Surrounding Mining: Las Bambas, Peru” which used the Grey Clustering Method in order to evaluate the impact in the area that includes the Challhuahuacho and Ferrobamba rivers where the las Bambas mine operate [13]. This information is important to establish the relationship between mining and water pollution, which is decisive because in the present study case the Crisnejas River is close to mining operations [14].

In the article “Water Quality Assessment of the Mining-Impacted Elqui River Basin, Chile” [15], the water quality assessment is done by using multivariate data analysis to characterize the main impacts (mining, agriculture and hydrothermal pollution) on Elqui River in Chile. Also, the use of factorial indices like mining pollution and salinization help to highlight and identify the sections of the river that were more influenced. For the study, this will be useful to consider some statistical methods like principal component analysis, determine the threats for the Crisnejas River Basin.

According to the paper “Hydrochemical evaluation of the influences of mining activities on river water chemistry in central northern Mongolia” [16], different concentrations of studied parameters give information about the setting of potential environmental activities, like mining and erosion processes. In the present work, a similar situation is studied by analyzing a river basin that is connected at the top of the watershed with a mine, and downhill with agriculture and livestock zones.

Additionally, the article “Finding water quality trend patterns using time series clustering: A case study” [17], explains the use of time series clustering to find time quality trend patterns in Zhejiang Province, China. In that way, there can be analyzed geographically distant regions, which may present similar patterns according to certain physicochemical parameters. As a result, finding root causes of water pollution, by anthropogenic factors would be more identifiable.

III. METHODOLOGY

Grey System theory is a methodology used in studies with small samples or lack of information. The grey clustering evaluation method based on CCTWF is used in evaluation of water qualities as used in previous studies [18].

A. First Step: Setting of Central Points

The central points are calculated by using a standard rule for water as delimitation points. Consequently, the need to convert the ranges into three Grey Classes (λ_1 , λ_2 and λ_3) used

in Peruvian regulation, makes important to calculate central points as averages and limits.

B. Second Step: Nondimensional Conversion

Converting original values to non-dimensional values is necessary for both standards and monitoring points. Then, they follow a matrix arrangement: $Z = \{Z_{ij}; i = 1, 2, \dots, m; j = 1, 2, \dots, n\}$, and for each criterion C_j ($j = 1, 2, \dots, n$). The nondimensional conversion is calculated using (1).

$$P_{ij} = \frac{z_{ij}}{\frac{\sum_{j=1}^n z_{ij}}{n}} \quad (1)$$

C. Third Step: Set the Grey Functions or Triangular Functions

The functions will be defined under the parameters established in previous step using Grey System functions (as shown in Fig. 1).

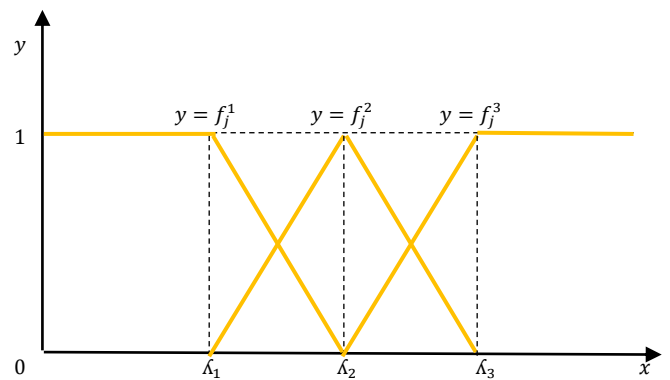


Fig. 1. CTWF Representation.

According to the categories assigned to each function, those established in the Peruvian DS were applied, which are the following:

$y = f_j^1 = A1$ = water for unrestricted vegetable irrigation

$y = f_j^2 = A2$ = water for restricted vegetable irrigation

$y = f_j^3 = A3$ = water for animal drinking

Under these conditions the functions are as shown in (2) – (4).

$$f_j^1(x_{ij}) = \begin{cases} 0, x \in [\lambda_j^2, +\infty > \\ \frac{\lambda_j^2 - x}{\lambda_j^2 - \lambda_j^1}, x \in < \lambda_j^1, \lambda_j^2 > \\ 1, x \in [0, \lambda_j^1] \end{cases} \quad (2)$$

$$f_j^2(x_{ij}) = \begin{cases} 0, x \in [0, \lambda_j^1] U [\lambda_j^3 + \infty > \\ \frac{\lambda_j^3 - x}{\lambda_j^3 - \lambda_j^2}, x \in < \lambda_j^2, \lambda_j^3 > \\ \frac{x - \lambda_j^1}{\lambda_j^2 - \lambda_j^1}, x \in < \lambda_j^1, \lambda_j^2 > \end{cases} \quad (3)$$

$$f_j^3(x_{ij}) = \begin{cases} 0, x \in [0, \lambda_j^2] \\ 1, x \in [\lambda_j^3, +\infty) \\ \frac{x - \lambda_j^2}{\lambda_j^3 - \lambda_j^2}, x \in [\lambda_j^2, \lambda_j^3) \end{cases} \quad (4)$$

D. Fourth Step: Determination of the Weight for each Criterion

In this step the clustering weight of the grey class parameters will be determined using the harmonic mean method expressed in (5).

$$n_j^k = \frac{\frac{1}{\lambda_j^k}}{\sum_{j=1}^n \frac{1}{\lambda_j^k}} \quad (5)$$

E. Fifth Step: Determination of the Clustering Coefficient

Next step is to calculate the clustering coefficient for each monitoring point $i, i = 1, 2, \dots, m$, using (6), and their corresponding grey class $k, k = 1, 2, 3$.

$$\sigma_i^k = \sum_{j=1}^n f_j^k(x_{ij}) \cdot n_j \quad (6)$$

Where $f_j^k(x_{ij})$ the value from the CTWF and n_j is the weight for each parameter.

F. Sixth Step: Determination of the Max Coefficient

Finally, to determine the category for each monitoring point using the maximum value of coefficient, (7) will be applied.

$$\max_{1 \leq k < s} \{\sigma_i^k\} = \{\sigma_i^{k^*}\} \quad (7)$$

IV. CASE STUDY

The analysis of superficial water quality was carried out in the Cuenca Crisnejas - Subcuenca Cajamarquino - ALA Cajamarca, located at the northern part of Perú [7]. Therefore, the closeness and influence of the watershed with Minera Yanacocha Project is important to be analysed. Furthermore, three principal rivers were considered for the study: river Mashcon, river Chonta and river Cajamarquino.

A. Definition of Objects Study

The study conducted by the “ALA Cajamarca” and the “Área Técnica de la Autoridad Administrativa del Agua V1 Marañón”, considered thirty monitoring points along the watershed. However, for the study we took only fifteen monitoring points (as shown in Table I), as they were considered to be strategically located (see Fig. 2).

B. Definition of Evaluation Criteria

The evaluation criteria used for the present study is determined by water quality parameters according to the study made by “ALA Cajamarca” and the “Área Técnica de la Autoridad Administrativa del Agua V1 Marañón” (shown in Table II). In addition, the criterions selected are chemically correlated to the type of deposit of Minera Yanacocha, which is a High Sulphidation Epithermal gold deposit.

C. Definition of Grey Classes

The definition of Grey Classes was based on the criterions of ECA 2017 (Table III) [9]. The analysis was made for

Category 3: Watering of vegetables and drink for animals, because Minera Yanacocha treats its waters for agriculture and livestock. In this study it is taken into account that “not restricted” irrigation values are more rigorous than the “restricted” irrigation water, and even more than “drink for animals”. Therefore, each category corresponds to λ_1, λ_2 and λ_3 as it corresponds from the highest, towards the lowest water quality.

TABLE I. SELECTED MONITORING POINTS

Point	Code	Coordinates (UTM WGS 84)	
		East	North
1	RQuil1	766920	9220901
2	RQuil2	767783	9216458
3	RRonq2	772043	9208237
4	RMash1	773157	9212710
5	RMash2	778523	9207039
6	RPorc1	771112	9214183
7	LMama1	796178	9226257
8	MVent1	787251	9222367
9	RChon1	787130	9216458
10	QHier1	789525	9227548
11	QArna1	780689	9227007
12	QChaq1	780760	9224459
13	RCaja1	798143	9193896
14	RQuin1	797025	9213249
15	QChul1	796890	9213344

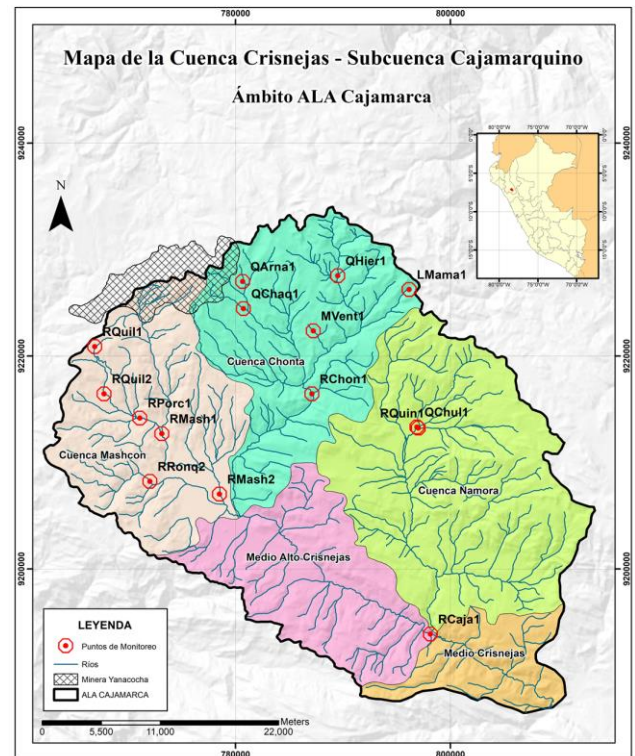


Fig. 2. Map of Study.

TABLE II. NOTATION OF CRITERIONS

Criterion	Units	Notation
OD	mg O ₂ /L	C1
pH	1-14	C2
DBO ₅	mg/L	C3
DQO	mg/L	C4
Al	mg/L.	C5
Fe	mg/L	C6
Mn	mg/L	C7
Thermotolerant Coliforms	NMP/100mL	C8
Esterichia Coli	NMP/100mL	C9
As	mg/L	C10
Pb	mg/L	C11

TABLE III. ORIGINAL CRITERIONS FOR GREY CLASSES

Parameters	Quality Index Condition		
	Irrigation of vegetables not restricted	Irrigation of vegetables restricted	Animal Drink
OD (mg/L)	≥ 4		≥ 5
pH (1-14)	6.5 - 8.5		8.4
DBO ₅ (mg/L)	15		15
DQO (mg/L)	40		40
Al (mg/L)	5		5
Fe (mg/L)	5		-
Mn (mg/L)	0.2		0.2
Thermotolerant Coliforms (NMP/100mL)	1000	2000	1000
Esterichia Coli (NMP/100mL)	1000	-	-
As (mg/L)	0.1		0.2
Pb (mg/L)	0.05		0.05

D. Calculations by using CTWF Method

1) *First step*: Setting of central points: In this step, some adjustments are necessary to be made, to create standards for diverse classes (Table IV). For example, take the higher and lower value to calculate the average as a medium index.

2) *Second step*: Non – dimensional conversion:

- **Standard Values**: For each criterion, and considering the three Grey Classes, it is calculated an average. Then, the new nondimensional standard value is the result of dividing the original value by the average (Table V).
- **Monitoring Points**: Using each average calculated above, the new non-dimensional monitoring point value, is the result of dividing the original value by the average (Table VI).

3) *Third step*: The values (λ₁, λ₂, λ₃) of each type of variable (pH, OD and more) are substituted in (8) – (10) in

order to obtain the functions that will be used to evaluate all the monitoring points. It is presented the functions of the dissolved oxygen parameter (OD) and in the same way the other parameters will be developed.

$$f_{C1}^1 = \begin{cases} 1, x \in [0, 0.889] \\ \frac{1-x}{1-0.889}, x \in < 0.889, 1 > \\ 0, x \in [1, +\infty > \end{cases} \quad (8)$$

$$f_{C1}^2 = \begin{cases} \frac{x-0.889}{1-0.889}, x \in < 0.889, 1 > \\ \frac{1.111-x}{1.111-1}, x \in [1, 1.111 > \\ 0, x \in [0,0.889] \cup [1.111, +\infty > \end{cases} \quad (9)$$

$$f_{C1}^3 = \begin{cases} \frac{x-1}{1.111-1}, x \in < 1, 1.111 > \\ 1, x \in [1.111, +\infty > \\ 0, x \in [0, 1] \end{cases} \quad (10)$$

TABLE IV. MODIFIED CRITERIONS FOR GREY CLASSES

Parameters	Quality Index Condition		
	Irrigation of vegetables not restricted	Irrigation of vegetables restricted	Animal Drink
C1	4	4.5	5
C2	6.5	7.45	8.4
C3	10	12.5	15
C4	30	35	40
C5	3	4	5
C6	3	4	5
C7	0.1	0.15	0.2
C8	1000	1500	2000
C9	500	750	1000
C10	0.1	0.15	0.2
C11	0.003	0.004	0.005

TABLE V. NON – DIMENSIONAL STANDARD VALUES

Codes	Quality Index Condition		
	λ ₁	λ ₂	λ ₃
C1	0.889	1.000	1.111
C2	0.872	1.000	1.128
C3	0.800	1.000	1.200
C4	0.857	1.000	1.143
C5	0.750	1.000	1.250
C6	0.750	1.000	1.250
C7	0.667	1.000	1.333
C8	0.667	1.000	1.333
C9	0.667	1.000	1.333
C10	0.667	1.000	1.333
C11	0.750	1.000	1.250

TABLE VI. NON – DIMENSIONAL MONITORING POINTS VALUES

Code	Quality Index Condition										
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
G1	1.538	0.787	0.160	1.114	0.033	0.053	0.130	0.001	0.002	0.000	0.050
G2	1.607	1.000	0.160	0.486	0.499	0.487	1.924	2.333	2.267	0.007	0.250
G3	1.591	1.164	0.160	0.286	0.117	0.101	0.130	0.307	0.227	0.003	0.100
G4	1.669	1.050	0.160	0.057	0.424	0.323	1.029	0.733	0.440	0.006	0.275
G5	1.396	1.126	1.360	2.486	3.178	3.365	3.769	73.333	9.333	0.081	0.710
G6	1.573	1.066	0.160	0.514	1.035	0.845	1.967	0.867	0.653	0.009	0.375
G7	1.218	0.728	0.160	0.000	0.033	0.293	0.104	0.001	0.002	0.014	0.050
G8	1.060	1.031	0.160	0.086	0.001	0.000	0.000	0.001	0.003	0.000	0.05
G9	1.616	1.114	0.160	0.286	0.156	0.152	0.392	1.133	1.467	0.011	0.175
G10	1.551	1.150	0.160	1.457	0.015	0.024	0.075	0.042	0.061	0.008	0.050
G11	1.460	0.580	0.320	0.571	0.276	0.456	0.066	0.003	0.003	0.051	0.050
G12	1.549	0.830	0.400	0.200	0.162	0.078	0.241	0.001	0.002	0.038	0.600
G13	1.549	1.111	0.160	0.343	0.757	0.721	1.347	7.333	0.009	0.017	0.550
G14	1.633	1.153	0.160	0.514	0.050	0.056	0.118	0.047	0.061	0.004	0.050
G15	1.551	1.122	0.160	0.286	0.052	0.064	0.062	0.187	0.293	0.000	0.050

The following tables present the evaluation of the parameters in the monitoring points established at the beginning. In Table VII, the non-dimensional values are shown.

4) *Fourth step:* As already mentioned in the methodology, below (Table VIII) are the weights for each parameter using the harmonic mean method through (5).

5) *Fifth step:* The clustering coefficient was calculated using (6). Table IX shows the results for each parameter.

6) *Sixth Step:* Finally, the highest value of the clustering coefficients for each point is chosen, then it determines the grey class which it belongs (7). For each point, the results are shown in Table X.

TABLE VII. PARAMETERS EVALUATED AT ALL MONITORING POINTS

Code	Eqs.	C1	C2	C3	C4	C5	C6
RQuil1	f1 (x)	0	1	1	0	1	1
	f2 (x)	0	0	0	0.20	0	0
	f3 (x)	1	0	0	0.79	0	0
RQuil2	f1 (x)	0	0	1	1	1	1
	f2 (x)	0	1	0	0	0	0
	f3 (x)	1	0	0	0	0	0
RRonq2	f1 (x)	0	0	1	1	1	1
	f2 (x)	0	0	0	0	0	0
	f3 (x)	0	1	0	0	0	0
RMash1	f1 (x)	0	0	1	1	1	1
	f2 (x)	0	0.61	0	0	0	0
	f3 (x)	1	0.39	0	0	0	0
RMash2	f1 (x)	0	0	0	0	0	0
	f2 (x)	0	0.02	0	0	0	0
	f3 (x)	1	0.98	1	1	1	1

TABLE VIII. WEIGHTS

Weight	λ_1	λ_2	λ_3
C1	0.077	0.091	0.101
C2	0.078	0.091	0.100
C3	0.085	0.091	0.094
C4	0.079	0.091	0.098
C5	0.091	0.091	0.090
C6	0.091	0.091	0.090
C7	0.102	0.091	0.084
C8	0.102	0.091	0.084
C9	0.102	0.091	0.084
C10	0.102	0.091	0.084
C11	0.091	0.091	0.090

TABLE IX. CLUSTERING COEFFICIENT

Points	λ_1	λ_2	λ_3
G1	0.844	0.018	0.180
G2	0.539	0.091	0.354
G3	0.818	0.000	0.100
G4	0.723	0.156	0.147
G5	0.102	0.001	0.914
G6	0.557	0.211	0.250
G7	0.844	0.000	0.101
G8	0.845	0.111	0.079
G9	0.641	0.064	0.308
G10	0.766	0.000	0.299
G11	0.844	0.000	0.200
G12	0.923	0.000	0.101
G13	0.639	0.015	0.356
G14	0.845	0.000	0.201
G15	0.845	0.004	0.196

TABLE X. VALUE OF MAX. CLUSTERING COEFFICIENT

Points	Max-Cof	Category
G1	0.8439	λ_1
G2	0.5389	λ_1
G3	0.8181	λ_1
G4	0.7230	λ_1
G5	0.9140	λ_3
G6	0.5565	λ_1
G7	0.8439	λ_1
G8	0.8453	λ_1
G9	0.6411	λ_1
G10	0.7659	λ_1
G11	0.8439	λ_1
G12	0.9234	λ_1
G13	0.6385	λ_1
G14	0.8453	λ_1
G15	0.8453	λ_1

V. RESULTS AND DISCUSSION

A. About the Case Study

In this report, grey clustering was used to classify the different monitoring points taken by ANA during 5 to 10 April, 2019. They were classified according to Peruvian legislation in the ECA – Category 3, “Irrigation of vegetables and animal drink” which is divided from more towards less rigorous: A1, A2 and A3.

Table X indicates the classification of each monitoring according to the category assigned. It is observed that 14 points belong to the category of water for “unrestricted irrigation”, which allows its use for irrigation of food crops that can be in direct contact with water and that can be consumed raw. On the other hand, one monitoring point (G5) belongs to “animal drinking” waters used for drinking by large animals such as cattle, horses or camelids, and for smaller animals such as pigs, sheep, goats, guinea pigs, birds and rabbits.

Fig. 3 shows that most of the monitoring points, even the ones that are close to the mines, present good water quality with few physicochemical variations, pH, etc. However, further downstream the point G5 is a monitoring point from Mashcon River located near to an urbanized area, and has low water quality because of the impact of direct sewage discharges. Besides, it can be verified by observing its high COD value [19].

B. About the Methodology

By way of comparison, the method used considers the uncertainty in its analysis unlike other methods that do not include it in their development. For example: Delphi or the Analytic Hierarchy Process (AHP). This is considered important in the study topic since the degrees of evaluation develop accuracy and decrease the uncertainty.

One of the advantages of using grey clustering is the simplicity of mathematical modeling, as beyond the results are more simple to understand the entire process (using the CTWF) [20]. In addition, the application of weights by means of the harmonic mean method makes the study more straightforward and uncomplicated. Additionally, the method is very useful, since it allows to classify and weight the data of the sample.

On the other hand, the development of the principal issue is subjected to the legislation of the country (which may or may not be well defined), in which was determined to carry out the study. And that is why the variability presented by this method is disadvantageous if it is to be compared with other studies.

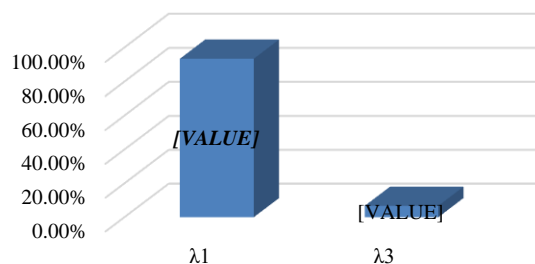


Fig. 3. Comparison of Monitoring Points according to their Categories.

Finally, the subject of study develops real problems that in its absence are always altered or are very changeable within the environment. That is where decisions will have to be made about negative anomalies detected under a dynamic process over time [21]. However, Grey Clustering is an effective methodology for environmental impact studies on surface water qualities [22][23].

VI. CONCLUSION

From the results obtained in the present study, it can be determined that in general there is good water quality in all areas, except in one point. This monitoring point is located in the urbanized area and is polluted due to sewage. On the other hand, although monitoring points related to the Yanacocha mine were taken into account, the results show that they are qualified waters that can be used for food consumption that is consumed raw.

With regard to the method used for this research, the Grey Clustering Method is one of the most effective options for classifying water monitoring points based on the information provided by the National Water Authority (ANA) and the parameters (ECA) established in Peruvian legislation. Unfortunately, the information only provides with values with no well established parameters and limits. Even though the samples taken lack of certainty, the method considers that flaw and mitigate it. On the other hand, the mathematical advantage of carrying out the study was seen, due to the fact that it is simple to apply triangular functions and to use the harmonic mean to establish the weights of the parameters. However, it is not efficient to use the Excel software to carry out such calculations required by this method, that is why a programming software should be used or without going so far as to use the Visual Basic that comes included in Excel.

Although the results indicate that water resource management is being carried out efficiently, with this study only random points belonging to the Crisnejas River Basin have been monitored. For this reason, it is proposed to carry out other investigations taking into account the discharge points of the Yanacocha mine in order to verify that the maximum permissible discharge limits are being accomplished. This is suggested because there is a possibility of dilution of pollutants by other effluents that when they converge with effluents directly affected by the Yanacocha mine; do not really reflect the problematic of water quality.

In addition, other studies of this area could be carried out using other methods similar to grey clustering in order to obtain another perspective and corroborate the results.

REFERENCES

- [1] A. P. Deditius, S. Utsunomiya, P. Sanchez-Alfaro, M. Reich, R. C. Ewing, and S. E. Kesler, “Constraints on Hf and Zr mobility in high-sulfidation epithermal systems: formation of kosnarite, KZr₂(PO₄)₃, in the Chaquicocha gold deposit, Yanacocha district, Peru,” *Miner. Depos.*, vol. 50, no. 4, pp. 429–436, 2015, doi: 10.1007/s00126-015-0586-z.
- [2] N. R. Haddaway et al., “Evidence of the impacts of metal mining and the effectiveness of mining mitigation measures on social-ecological systems in Arctic and boreal regions: A systematic map protocol,” *Environ. Evid.*, vol. 8, no. 1, pp. 1–11, 2019, doi: 10.1186/s13750-019-0152-8.
- [3] A. Delgado and H. Flor, “Selection of the best air purifier system to urban houses using AHP,” in 2017 CHILEAN Conference on Electrical,

- Electronics Engineering, Information and Communication Technologies, CHILECON 2017 - Proceedings, 2017, vol. 2017-Janua, doi: 10.1109/DISTRA.2017.8229622.
- [4] A. Delgado and I. Romero, "Environmental conflict analysis on a hydrocarbon exploration project using the Shannon entropy," in Proceedings of the 2017 Electronic Congress, E-CON UNI 2017, 2018, vol. 2018-Janua, doi: 10.1109/ECON.2017.8247309.
- [5] A. Delgado, D. Vriclizar, and E. Medina, "Artificial intelligence model based on grey systems to assess water quality from Santa river watershed," in 2017 Electronic Congress (E-CON UNI), Nov. 2017, pp. 1–4, doi: 10.1109/ECON.2017.8247310.
- [6] A. Delgado, P. Montellanos, and J. Llave, "Air quality level assessment in Lima city using the grey clustering method," Jan. 2019, doi: 10.1109/ICA-ACCA.2018.8609699.
- [7] Ministerio de Agricultura and Autoridad Nacional del Agua, "Octavo Monitoreo Participativo de Calidad de Agua Superficial de la Cuenca del Río Crisnejas - Sub Cuenca Cajamarquino," 2019.
- [8] C. Aubron, H. Cochet, G. Brunschwig, and C. H. Moulin, "Labor and its productivity in andean dairy farming systems: A comparative approach," *Hum. Ecol.*, vol. 37, no. 4, pp. 407–419, 2009, doi: 10.1007/s10745-009-9267-9.
- [9] Ministerio del Ambiente, "Estándares de Calidad Ambiental para Agua (ECA)," El Peru., pp. 6–9, 2017.
- [10] S. Liu and Y. Lin, *Grey Systems*, vol. 68. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
- [11] C. T. Son, N. T. H. Giang, T. P. Thao, N. H. Nui, N. T. Lam, and V. H. Cong, "Assessment of Cau River water quality assessment using a combination of water quality and pollution indices," *J. Water Supply Res. Technol.*, vol. 69, no. 2, pp. 160–172, Mar. 2020, doi: 10.2166/aqua.2020.122.
- [12] D. Szpak and B. Tchórzewska-Cieślak, "The Use of Grey Systems Theory to Analyze the Water Supply Systems Safety," *Water Resour. Manag.*, vol. 33, no. 12, pp. 4141–4155, 2019, doi: 10.1007/s11269-019-02348-y.
- [13] V. Bax, W. Francesconi, and A. Delgado, "Land-use conflicts between biodiversity conservation and extractive industries in the Peruvian Andes," *J. Environ. Manage.*, vol. 232, pp. 1028–1036, Feb. 2019, doi: 10.1016/j.jenvman.2018.12.016.
- [14] A. Delgado, A. Espinoza, P. Quispe, P. Valverde, and C. Carbajal, "Water quality in areas surrounding mining: Las Bambas, Peru," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 4427–4432, Oct. 2019, doi: 10.35940/ijitee.L3807.1081219.
- [15] L. Ribeiro et al., "Water Quality Assessment of the Mining-Impacted Elqui River Basin, Chile," *Mine Water Environ.*, vol. 33, no. 2, pp. 165–176, 2014, doi: 10.1007/s10230-014-0276-6.
- [16] B. Batsaikhan et al., "Hydrochemical evaluation of the influences of mining activities on river water chemistry in central northern Mongolia," *Environ. Sci. Pollut. Res.*, vol. 24, no. 2, pp. 2019–2034, 2017, doi: 10.1007/s11356-016-7895-3.
- [17] L. Huang, H. Feng, and Y. Le, "Finding water quality trend patterns using time series clustering: A case study," *Proc. - 2019 IEEE 4th Int. Conf. Data Sci. Cyberspace, DSC 2019*, pp. 330–337, 2019, doi: 10.1109/DSC.2019.00057.
- [18] L. N. Zhang, F. P. Wu, and P. Jia, "Grey Evaluation Model Based on Reformative Triangular Whitenization Weight Function and Its Application in Water Rights Allocation System," *Open Cybern. Syst. J.*, vol. 7, no. 1, pp. 1–10, 2013.
- [19] D. Mercado-Garcia et al., "Assessing the freshwater quality of a large-scale mining watershed: The need for integrated approaches," *Water (Switzerland)*, vol. 11, no. 9, pp. 1–20, 2019, doi: 10.3390/w11091797.
- [20] M. Sacasqui, J. Luyo, and A. Delgado, "A Unified Index for Power Quality Assessment in Distributed Generation Systems Using Grey Clustering and Entropy Weight," Dec. 2018, doi: 10.1109/ANDESCON.2018.8564631.
- [21] Y. Liu and R. S. Zhang, "A three-way grey incidence clustering approach with changing decision objects," *Comput. Ind. Eng.*, vol. 137, p. 106087, Nov. 2019, doi: 10.1016/j.cie.2019.106087.
- [22] X. Q. Fu and Z. H. Zou, "Water Quality Evaluation of the Yellow River Basin Based on Gray Clustering Method," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 128, no. 1, 2018, doi: 10.1088/1755-1315/128/1/012139.
- [23] J. Wang et al., "Application of Grey Clustering Method Based on Improved Analytic Hierarchy Process in Water Quality Evaluation," *MATEC Web Conf.*, vol. 246, pp. 3–7, 2018, doi: 10.1051/mateconf/201824602004.

Analysis of Speech Signal Data of Missing Vowels using Logistic Regression and K-Means Clustering

Ujjal Saikia¹

Centre for Computer Science and Applications
Dibrugarh University, Dibrugarh, India

Jiten Hazarika²

Department of Statistics
Dibrugarh University, Dibrugarh, India

Abstract—In this paper, an attempt has been made to study and analyze speech signal data. Here, the sound or speech data has different attributes like time, pitch, formant frequencies, speaker type, Vowel No etc. The dataset used here is speech signal data which are analog in nature and has been converted to digital format. After converting the data into digital format we want to establish a Logit model to predict the speaker gender on the basis of the pitch signal values which is also considered as fundamental formant frequency. That is our objective is to predict whether a speaker is male or female by looking at the pitch value by using logistic regression. We have applied clustering techniques to visualize and interpret how it works in speech signal data. The logistic model gives us 91% accuracy rate with low and efficient AIC value where as in case of the clustering algorithm we get a 93% accuracy for the whole sample.

Keywords—Clustering methods; formant frequency; Logit model; pitch; SType

I. INTRODUCTION

The role of statistical techniques in data analysis is vital. Proper use of statistical techniques in data analysis can provide very fruitful result. The key thing is that we have to select proper methodologies and techniques. One of the most frequently used method is the Regression analysis. The regression analysis can give us a very good idea about the whole dataset while applied carefully and selectively. Regression methods like simple linear regression, multiple linear regression are frequently used for data analysis where the relationship between dependent and independent variables are linear in nature. While Logistic regression is preferred when the regressand is a categorical variable. Through the literature Survey it has been observed that various statistical methods like Regression, Classification etc. have been applied in different types of data. Greenstein, J. [1954], shows the effect of television viewing upon elementary school grades using multiple and partial linear regression analysis in educational dataset [1]. Warner and Misra [1996] discussed the use of neural network and compares it with traditional regression analysis. They show that neural network works as a nonparametric regression model and it enables to model complex functional form. They also discussed the advantages and difficulties of using neural network against the use of regression analysis [2].

Gibbs et al. [2006] shows the use of regression analysis to establish the relationship between home environment and reading achievement for institutional cum educational datasets

[3]. Pao[2008] made an empirical study on comparison of neural network and multiple regression analysis in modelling capital structure. This study adopted multiple linear regressions and artificial neural networks models with seven explanatory variables of corporation's feature and three external macro-economic control variables to analyse the important determinants of Capital Structure data [4]. Keshavarzi & Sarmadian[2010] compared Artificial Neural Network and Multivariate Regression in Prediction of Soil cation exchange capacity. Investigation of soil properties like Cation Exchange Capacity (CEC) plays important roles in study of environmental research as the spatial and temporal variability of this property have been led to development of indirect methods in soil data analysis [5]. Prica and Sinisa[2010] has done experimental study for recognizing Vowels in continuous speech by using parameters like formant frequency in speech signal data [6]. Ganesan et al. [2010] shows the Application of statistical and machine learning techniques in Diagnosing Cancer Disease using Demographic Data [7]. Raghavendra and Srivatsa[2011] evaluated Logistic Regression and Neural Network Model with Sensitivity Analysis on medical datasets. The goal of this research work was to compare the performance of logistic regression and neural network models on publicly available medical datasets [8]. Al-Shayea [2011] also shows the use of ANN modelling in Medical Diagnosis [9]. From the experimental results, it is confirmed that the neural network model with sensitivity analysis gives more efficient result. Different visual plotting techniques has already been used in speech signal data the visually detect the significance of the dataset and its different attributes. Rehman and Hazarika [2014] have done experimental studies for analyzing and recognition of vowels of low resource languages by using speech signal dataset [10]. Another relevant study of using statistical technique is the use of it in the verification and identification of speaker. Zhang [2018], has shown the use of Linear Regression for Speaker Verification [11]. There is also scope in using Statistical methods while studying Acoustic properties of speech signal data. Some previous studies pointed out how to apply basic statistical methods. Saikia et al. [2019] has used in effective data visualization [12]. Jiang et al. [2020] has described statistical methods for Feature Extraction Method for Speaker Recognition very efficiently in their work with proven experimental outcomes and results [13]. Magdiel and Pilar [2021] used standardized domain adaptation techniques for classification in imagined speech recognition [14]. Babak et al. [2021] reviews deep learning approaches in speech emotion recognition by using machine learning techniques.

They have applied these techniques in already available statistical datasets [15].

In our present study we have applied statistical as well as artificial machine learning methods in sound data. We have given our emphasis on building regression model in speech signals with the following objectives.

- To build a model by applying logistic regression to identify male and female speaker on the basis of the pitch values.
- To apply clustering methods to the speech signal data to categorize male and female speakers separately and check the efficiency of the model.

II. BACKGROUND AND METHODOLOGIES

A. Speech Signal Data

Speech signal data are presentation of sound data in digital format. It is basically an analogue sound recorded in a closed environment. Closed or sound proof environment is a necessary condition to record sound for analytical purposes. Otherwise it may be affected by other sounds. Noisy environment may hamper to get the actual sound regarding the pronunciation of a particular vowel or any words which is to be prepared for analysis. A sound signal data after converting to digital dataset may have the following attributes. Let us give a brief idea about some of the attributes and parameters of the sound data.

1) *Time*: It is the duration or time period for which a sound is uttered by human generally represented in seconds during vowel utterance.

2) *Speaker type*: It is the type of the speaker. For example whether the sound is pronounce by male or female. Or a speaker from native or non-native background etc. This variable is categorical in nature.

3) *Pitch (F0)*: It is known as fundamental frequency. The pitch is defined as the rate of vibration of the vocal chords of the person who is pronouncing the particular sound. This pitch frequency varies across different sounds. It also varies depending on the speaker type.

4) *Formant frequencies*: One of the distinguishing frequency component of human speeches are the formant frequencies. These values are also obtained from the recorded sound. It is considered as the particular resonance frequency of the vocal tract which is considered to have the maximum frequency during vowel utterance. The formant frequencies are denoted by the symbols F1, F2 etc.

5) *Vowel no*: It refers to the serial numbers of the Vowels of the language while considering for analysis. Here in Mising language we have a total of 14 Vowels.

B. About Mising Language

Here in our study we have selected sound data which is related to vowel utterance of Mising language. So it is convenient to discuss about the language "Mising". It is a language from North-Eastern India. It is a mixture of Indo-Aryan and Sino-Tibetan family of language. This language is

considered as linguistic offshoot of the Tibeto-Burman branch of Sino-Tibetan family. According to Census 2001 only 5 Lakhs people left who speaks this language and day by day it is decreasing consistently. The people who can write this language by using their own script is also very less. It is considered as one of the low resource languages that are spoken by people in Assam. The language is considered as low resource language because of lack of content in internet and online resource. The Mising language has 14 vowels and 15 consonant. In our present study we have included the speech signals of the following 14 vowels of Mising language. They are: /i/ , /i:/, /e/ , /e:/, /a/ , /a:/, /o/ , /o:/ , /u/ , /u:/ , /é/ , /é:/ , /í/ , /í:/.

C. Regression Methods

Regression Analysis is one of the most widely used data analysis technique for prediction. It is used to investigate the relationship between dependent and independent variables. The dependent variable is known as target variable while the independent variable are known as predictor variable. The terms "Regressand" and "Regressor" are also used for dependent and independent variables respectively under study. Most common types of regression techniques are outlined below.

1) *Simple linear regression*: This method is adopted while we have one target variable and one predictor variable. The simple linear regression is the simplest of all the others and can be successfully implemented in different situations in case of two variables when the relationship between them is linear in nature.

2) *Multiple linear regression*: Multiple linear regression is also known as multiple regression. This method is used when we have to predict the outcome of a dependent variable using several independent or explanatory variables. This method is used in case of more than two variables. The relationship among them is assumed to be linear in nature.

3) *Logistic regression*: Logistic Regression is a regression model to model a binary dependent variable. The model is known as logistic or logit model and it is used to model the probability of a certain class or event such as Yes/No, Pass/Fail, Male/Female, Affected/Not affected etc. That is, this technique is used when regressand is basically a categorical variable. Here in our current study we have tried to build a logit model for Speaker type when we are given the pitch value. We also determined model accuracy on the basis of true prediction and false prediction for the total number of test data on the basis of the train model. We have checked the model with different parameters like Combination of F0 and VowelNo, F1, Vowel No etc. But all the models based on them is results in higher AIC values. AIC stands for the Akaike information criterion which is an estimator of out-of-sample prediction error, given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. Generally a lower AIC value is expected. Lower the Value of AIC better is the Model. After that we have checked

the same with some clustering algorithm which are based on machine learning.

4) *ANN models*: ANN stands for Artificial Neural Network. These models are useful in case of building nonlinear and complex relationships among variables under study. The models based on Artificial Neural Network may be used to model complex relationships between inputs and outputs that is for the independent and dependent variables or to find patterns in data and future prediction. A part of the dataset is trained for prediction and on the basis of the trained data outputs are checked for efficient prediction. ANN models are often considered as a computational or mathematical model which inspired by the functioning human brain that is biological neural networks.

5) *Clustering techniques*: Like regression analysis Cluster Analysis also very much widely used techniques for data analysis. The method is simply known as Clustering. Clustering is defined as a task of grouping of objects or data in such a manner that the objects in the same group are more similar to the each other compared to the objects of the other groups or classes. Clustering is one of the main techniques used in data mining or fact finding in data. It is a common technique in statistical practice which is widely used and applied in various fields like pattern recognition and machine learning. Here in our dataset we have applied K-means clustering method to group our data into different categories. Let us first give a brief idea about K-means Clustering.

6) *K-Means clustering*: K-means Clustering is one of the most commonly practiced unsupervised machine learning algorithm which is used for partitioning the whole data set into K no. of groups or cluster. In general K is specified prior to

the analysis by the data analyst. It classifies the objects or values into two or more groups in such a way that the objects within the same group or clusters are quite similar while the objects from the different clusters are quite dissimilar. Each cluster in K-means clustering is represented by its centroid or centre which corresponds to the mean of the points. The basic idea behind the K-means clustering is discussed below.

In K-means Clustering the total within cluster variation is minimized. There are many K-means algorithms. Here we have used the standard algorithm known as Hartigan-Wong algorithm (1979). This algorithm defines the Euclidean distances between item values and the corresponding centroid.

III. ABOUT THE DATASET

The dataset used here is secondary in nature while collected in analogue format. It has later been converted to digital format using the application software PRATT for analysis purpose. The dataset looks like as given in Table 1. The table 1 indicates the first 28 samples. The data recorded in a noise free environment inside lab and there after converted to digital format by using PRAAT Software.

In the next page we are showing the sample dataset with basic discussion and about pre-processed data. The attributes like Time, F0, F1, F2 Speaker Types or Speaker Gender (SType) and serial no of Vowels (VowelNo).

1) *Discussion about the data*: The Missing Data in Table 2 stands for digital dataset which are originally sound signals recorded in the laboratory. That is the data were originally of analogue signals which are converted to digital format by using PRAAT software in computer labs.

TABLE I. DATA DICTIONARY

Serial No.	Parameters and Detailed Info
1	The population size 70
2	Sample dataset size 28
3	Total Speakers Male: 28 Total Speakers Female: 42
4	Recording Environment: Laboratory(Closed Noise Free)
5	Recording Equipment Device: PHILIPS Microphone with Noise Cancellation feature
6	Channel: Mono
7	Frequency of the Sampling: 22050 Hz
8	Software: PRAAT
9	Language Spoken: MISING
10	Speakers: Graduates and Post Graduate Students of from MISING community
11	Speaker Type: Native

TABLE II. SAMPLE DATA SET OF FIRST 28 DATA POINTS

The First 28 sample points of Speech signal data of the Low Resource Language MISING					
Time	F0	F1	F2	SType	VowelNo
0.345669	235.222489	993.258705	1533.717724	1	1
0.282823	237.936687	175.815388	1141.23666	1	2
0.329955	244.257104	648.340679	1213.0288	1	3
0.3928	243.528907	744.776094	1173.473009	1	4
0.34568	197.99159	817.4236	1810.99103	1	5
0.209974	250.077406	836.121161	1754.962313	1	6
0.370915	280.340615	305.037348	1463.17897	1	7
0.298526	274.983809	476.091377	1587.204281	1	8
0.298526	250.441551	396.317328	598.567316	1	9
0.282823	270.121203	532.275868	1729.586239	1	10
0.3928	254.759789	613.597921	1023.284447	1	11
0.23568	240.601283	633.00456	984.396661	1	12
0.361372	271.812947	439.597155	807.948917	1	13
0.267109	254.08066	358.673299	664.206643	1	14
0.408209	127.535411	792.494447	1256.335354	0	1
0.343753	122.173526	743.720154	1209.341717	0	2
0.365238	133.058353	524.762969	1925.038107	0	3
0.408209	143.028187	542.013213	1961.55233	0	4
0.408209	136.462751	514.975377	1411.257475	0	5
0.408209	136.462751	514.976142	1411.259634	0	6
0.451179	143.237093	311.220514	2267.531138	0	7
0.386723	172.180501	325.989119	2336.622676	0	8
0.429694	148.157406	372.217248	1519.504391	0	9
0.451179	166.545805	337.593758	1364.08963	0	10
0.429694	143.768601	452.017997	836.50649	0	11
0.386723	128.887188	428.04571	800.253539	0	12
0.429694	138.032437	319.960304	800.001534	0	13
0.365238	145.755593	372.810288	823.053677	0	14

IV. RESULTS

A. Results of Logistic Regression Method

For our current study we are using logistic regression for model building by considering SType as dependent variable and F0 (pitch) as independent variable. We have used R programming for analysis purpose and following observations were made in our train data set. SType '1' means the speaker is Female and SType '0' means speaker is Male. Results of fitted logistic model are shown in Table 3.

Mylogit is the logistic regression model build for predicting SType from F0 values. We have predicted the value of SType by using the model which gives fruitful result. We have shown the predicted value along with the graphical plot as shown in Table 4.

TABLE III. BASIC RESULTS OF FITTED LOGISTIC MODEL

(R codes for building the model)		
Mylogit <- glm(SType ~ F0, data = mydata, family = "binomial")		
Coefficients		
(Intercept)	F0	
-19.7270	0.1186	
Degrees of Freedom:		
Total	Residual	
57	56	
Null Deviance:	Residual Deviance	AIC
77.9	11.92	15.92

While applying the Hartigan-Wong algorithm on the overall dataset we have observed that out of 70 records two clusters were prepared of sizes 41 and 29. [TABLE 5]. A total of 5 records has been observed which was predicted incorrectly on the basis of the pitch values where 4 incorrect prediction in male data whereas only 1 incorrect prediction in the female voice sample as shown from the Fig. 2 and the earlier tables of reference. Here in case of the clustering algorithm we get a 93% accuracy rate even in the whole sample.

The Clustering Algorithm based on machine learning is slightly ahead in terms of true prediction whereas the Logit model also gives us almost very accurate result in terms of small sample set. The logistic regression helps us in determine the shape of the prediction curve [Fig. 1] and comparing all the other significant factors from the other attributes. In our study we have found out that F0 is the most significant factor which can be used to predict the gender of the speaker.

However for some of the pitch values were incapable of predicting the type of the speaker in both the models due to general tendency of a female speaker having a male like voice and Vice Versa. This phenomenon may occur due to some external factors like noisy environment, recording disturbances etc.

Finally in case of data instances and statistical significance the most vital factor is that as we have mentioned earlier Mising is a low resource language as well as very few speakers are left in our region. It was managed to collect a few samples regarding the vowel pronunciation. We have used K-Means Clustering as it is originally based on signal processing and useful for small samples.

Currently we are collecting more sample voices and planning to use methods like SVM and other ANN models which are supposed to be more reliable.

VI. CONCLUSION AND FUTURE SCOPE

In the present study we have applied statistical techniques like Logistic regression to analyze the sound data and obtained interesting results as discussed above. We have also applied some other clustering methods based on machine learning. We have compared the results with relative advantages and efficiencies.

It will be helpful in future for comparing result with other data mining techniques as well ANN model and carry forward this work in gender classification in speech signals data. Similarly these works can be carry forwarded for other problems like Speaker identification from native and non-native language, frequency prediction etc. for other low resource languages as well.

ACKNOWLEDGMENT

The research work is based on a primary data set of speech signal data. We sincerely offer our thanks and gratitude to Dr. Rizwan Rehman, Assistant professor of Centre for Computer Science and Applications, Dibrugarh University for providing his data for analytical purposes.

We would also like to offer my sincere thanks to those students who have provided their voice samples for the community research work.

REFERENCES

- [1] Greenstein, J. (1954), Effect of television viewing upon elementary school grades, *The Journal of Educational Research*, 48, 161-176.
- [2] Warner, B. and M. Misra (1996), Understanding Neural Network as Statistical Tool, *The American Statistician* 50(4),pp. 284-293.
- [3] Gibbs Y. Kanyongo, Janine Certo, Brown, I.Launcelot, Using regression analysis to establish the relationship between home environment and reading achievement: A case of Zimbabwe, *International Education Journal*, 2006, 7(5), 632-641.
- [4] Pao,H.T, A Comparison of Neural Network and Multiple Regression Analysis in Modelling Capital Structure, *Expert Systems with Applications* 35 , 2008, pp.720-727.
- [5] Keshavarzi,A. and F. Sarmadian, Comparison of Artificial Neural Network and Multivariate Regression Methods in Prediction of Soil Cation Exchange Capacity, *International Journal of Environmental and Earth Sciences* , 2011(1),pp. 25-30.
- [6] B. Prica and Sinisa, Recognition of Vowels in Continuous Speech by Using Formants, *SER.: ELEC. ENER.* vol. 23, no. 3, December 2010, 379-393.
- [7] Ganesan,N., Venkatesh,K., Rama, M. A. and A.M.Palani[2010],Application of Neural Network in Diagnosing Cancer Disease using Demographic Data, *International Journal of Computer Applications* 1(26),pp.81-97.
- [8] Raghavendra B.K. & S.K. Srivatsa[2011], Evaluation of Logistic Regression and Neural Network Model With Sensitivity Analysis on Medical Datasets, *International Journal of Computer Science and Security* 5 (5) ,pp.503-511.
- [9] Al-Shayea,Q. K.[2011],Artificial Neural Networks in Medical Diagnosis, *International Journal of Computer Science Issues* 8(2),pp.150-154.
- [10] Rizwan Rehman and Gopal Chandra Hazarika, Analysis and Recognition of Vowels in SHAIYANG MIRI Language using Formants, *International Journal of Computer Applications* 89(2):7-10, March, 2014.
- [11] Xiao-Lei Zhang, Linear Regression for Speaker Verification, arXiv:1802.04113,v1, Pp.1-10, 2018.
- [12] Ujjal Saikia, Rizwan Rehman, Jiten Hazarika, Gopal Ch. Hazarika, Predictive Analysis Using Regression Methods in Low Resource Language "MISING", 2nd International Conference on information systems & management science (ISMS) 2019.
- [13] Jiang Lin ,Yi Yumei ,Zhang Maosheng ,Chen Defeng ,Wang Chao ,Wang Tonghan, A Multiscale Chaotic Feature Extraction Method for Speaker Recognition, *Complexity*, Hindawi, vol. 2020, pages 1-9, December.
- [14] Magdiel Jiménez-Guarneros and Pilar Gómez-Gil,, Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition, *Pattern Recognition Letters*, Volume 141, January 2021, Pages 54-60.
- [15] Babak Joze Abbaschian , Daniel Sierra-Sosa , Adel Elmaghraby, Deep Learning Techniques for Speech Emotion Recognition, *Sensors* 2021, 21(4),1-27.

Hybrid SFLA-UBS Algorithm for Optimal Resource Provisioning with Cost Management in Multi-cloud Computing

Muhammad Iftikhar Hussain¹, JingSha He², Nafei Zhu^{3*}
Fahad Sabah⁴, Zulfiqar Ali Zardari⁵, Saqib Hussain⁶, Fahad Razque⁷

Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China^{1, 2, 3, 4, 6, 7}
Department of Information & Communication Technologies
Begum Nusrat Bhutto Women University, Sukkur, Pakistan⁵

Abstract—Multi-cloud is a vendor-based heterogeneous cloud paradigm in recent era of computing with dynamic infrastructural deployment. Multi-cloud provides all the essential and on-demand requirements of a virtual environment from various domains under a single service level agreement (SLA). Consumers from multitier domains can access all the available resources placed in a shared pool on service provider's side, as per their requirement. The shared pool of resources creates complexity in assigning the best and suitable resource to a particular virtual instance under the same services provider end. The complexity of resources in terms of accessibility from the various domains, dynamic allocation, security, and quality of services (QoS) raises concerns in the multi-cloud infrastructure. This complexity raise concern relates to optimal provisioning and cost management. In the proposed work a hybrid technique with a shuffled leapfrog algorithm and ubiquitous binary search (SLFA-UBS) to resolve these issues with optimal provisioning, dynamic allocation and better resource selection. The proposed work will help to create a need-based and demand-based resource pool with the appropriate selection of each resource. The proposed model also supports resource optimization with dynamic provisioning, cost-effective solution to achieve QoS in multi-cloud deployment on service provider end.

Keywords—Optimal provisioning; resource allocation; multi-cloud; cost management; QoS and selection of resources

I. INTRODUCTION

Multi-Cloud is a blend of cloud paradigms, which provides various on demand services with dynamic and elastic resource allocation in a multitier environment. Multi-cloud offers pay per as you go business model for dynamical allocation of resources from a shared pool. Cloud services work under the signed service level agreement (SLA) and standard of architecture (SOA) between a cloud provider and a broker. Cloud computing has distributed nodes, and the end-user does not know about the structural changes and details of physical resources, allocated on-demand as per-use based on given SLA [1].

Dynamic Resource allocation strategy (RAS) applied on big clusters and data centers in multi-cloud infrastructure, where the pool of resources shared among multitier environment to cover the need base or on demand utilization. RAS provides elasticity of basic hardware, the basic hardware

part consists of Hard Disk, Memory, CPU and related I/O devices. These resources allocated or assigned dynamically on demand to achieve performance, QOS and cost management in multi-cloud architecture. Dynamic resource allocation maintains high-level security and resource provision to avoid the overload of resources and achieve the requirement of green computing with optimization [2], [3]. By using RAS, can be avoided the hotspot (less resources than demand) and cold spot (more resources than demand), also plays a vital role to control the other infrastructure-based parameters like efficacy and cost. RSA provide redundant backup link to survive in case of failure, which is a negative impact of infrastructure design, resources scheduling (by using resource provisioning algorithms) used to maintain backup and SLA requirements [4].

Dynamic nature of multi-cloud provides heterogeneity of resources, services, applications, and server in both Virtual Machines (VM) and Physical Machines (PM). Different algorithms used to merge VM and PM depend on the nature of the environment and dependency in terms of resource allocation from shared pool. These algorithms are based on probability or scheduling of resources. Isolation of resources and scalability achieved by using dynamic nature of multi-cloud with dynamic allocation and provisioning of resources.

In multi-cloud deployment where different cloud paradigms and services models combine under a single umbrella and shared pool of resources have heterogeneous requirement in terms of basic structure, deployment model, access rights, transparency and cost. In this architecture, there are lot of security and allocation concerns in terms of end user's rights transparency, shared pool of resources and services, dynamic and runtime allocation, cost management and efficacy. There are many pre-defined techniques for the allocation and optimization of the resources in hybrid cloud environment e.g. scheduling techniques, feature based allocation, priority allocation, clustering and scaling based algorithms. These techniques are useful to overcome dynamic allocation and provisioning in private or hybrid cloud model but these techniques have lack to overcome issues related to multi-cloud model, resources optimization and cost management. The proposed technique will provide a better way to overcome issues related to dynamic and on demand

*Corresponding Author

resource provisioning with optimization and cost management. The proposed model also helps to avoid security issue related to shared-pool of services in multi-cloud and helpful for the new user to figure out basic concerns to adopt multi-cloud services.

Multi-cloud is an archetype that supports vendors to provide services among various private and public clouds holding any blend of these domains, e.g., heterogeneous cloud vendors, accounts, application, services, deployment, security models, premises, regions, and availability zone. The proposed model designed a hybrid algorithm with a shuffled leapfrog algorithm (SLFA) and Ubiquitous binary search (UBS). The design algorithm is helpful for optimization, energy saving, and QoS. In this research deployed a unique algorithm, e.g., genetic algorithm (GA), SLFA, and UBS. The proposed design model shows better results in terms of efficiency and cost-saving model with QoS.

The research contributions are

- Define a new multi-cloud paradigm under IaaS to achieve optimal resource allocation.
- A hybrid algorithm to design the best selection and allocation of resources.
- Fulfill the need-based and demand-based requirements of a virtual environment in multi-cloud.
- Achieve cost-optimized and efficient solutions in a multi-cloud paradigm using SLFA-UBS.

This paper further contains three sections i) related work ii) proposed work iii) optimal hybridization iv) result and discussion v) conclusion.

II. RELATED WORK

Xiaoqun Yuan et al. [5] proposed a game theory that used math's strategies to draw the possible move and functions using Nash equilibrium. The accuracy of the activities makes more chances to win. Each player strategy is the best response for others. The proposed theory depends on memory channel scheduling concerning time. The resource allocation in proposed model assign by using geo distribution with optimal distribution. Mira Morcos et al [6] proposed a greedy algorithm and Nash equilibrium technique for best resource allocation with exponential time. Numerical analysis technique performed form mobile computing network with for resource allocation and maximum cost saving.

Seyedehmehrnaz Mireslami et al. [7] defined a dynamic resource allocation technique on uncertainty-based algorithms and optimization methods and saving total estimated cost. The uncertainty proposed model provides dynamic resource allocation with efficiency and accuracy in cloud infrastructure and deployment. Prasad Devarasetty [8] proposed need-based resource allocation to implement a genetics algorithm technique with a particular defined budget to achieve QoS. The designed techniques deployed on amazon based cloud with cost limitation and efficient resource allocation.

Lailan M. Haji et al. [9] described various techniques for dynamic resource allocation in a cloud environment to achieve

efficiency, accuracy, and QoS. J. Praveenchandar et al. [10] proposed an optimal model for dynamic resource allocation and cost-saving in a cloud environment with minimal power management. The authors used a prediction-based dynamic resource table-updating algorithm in the proposed solution to save optimal power. The proposed technique uses impressive task scheduling and power utilization techniques.

On-time deployment access control separation, direct and indirect trust between grid and cloud computing to avoid overlapping and secrecy of credentials, data, and resources. [11] Open stack used on Linux based Xen and KVM machines to extract there features on basic system level. System analysis engine SAE can investigate core resources like memory usage and optimization without any thread and handler. Discrete firefly algorithm applied to avoid side-channel attacks on shared resources to reduce malicious tenant access, energy consumption, and resource loss on the provider end (DFA-VMP) under IaaS [12]. All existing OS-based attacks work in the same way; separate functionally of content reduce risks [13].

Flora Amato et al. [14] stated data security validation and verification techniques using particular classification algorithms on databases. The thermal function used to classify software and hardware classification. Saurabh Singh et al. [15] defined three tire security models related to embed system architecture to achieve trust and VM migration from one source to another associated strategies.

Gururaj Ramachandra et al. [16] defined three attack vector networks, hardware, and hypervisor to retain the complexity of resource allocation and best utilization in a particular virtual system. S Javanmardi et al. [17] proposed application-scheduling parameters in the cloud with fuzzy logic, genetic algorithms, and clustering techniques to achieve QoS in efficiency and throughput. The proposed work helpful to meet the needs of resource pooling and allocation.

K. Dinesh Kumar et al [18] proposed a resource-provisioning model with cost saving strategies to avoid loss of resources. A perdition method presented to avoid over costing, energy lost in data center, and cloud based environment, the proposed method predicted the basic need and upcoming starvation of resources allocation and managed the load as per define perdition. The described techniques enhanced the accuracy, correlation and utilization based perdition of resources in cloud computing. Xiaolong Xu et al [19] presented a meteorological framework in cloud computing with resource provisioning, flat tolerance and load balancing. In the proposed framework, virtual layer 2 further extracted to define meteorological framework then a non-dominated sorting algorithm applied to get load balancing with flat tolerance. Xiaolong Xu et al [20] proposed an uncertainty-based software define framework form edge computing with balanced resource provision and cost in term of energy consumption. A multi-objective dynamic allocation with balanced scheduling techniques adopted to achieve energy efficient and cost effective dynamic resource provision and allocation in fog computing environment.

Cloud computing is multitier, highly scalable, and pay-per-use-based model over the internet; due to this nature, virus,

Trojans, and spy are the types of inherited attacks categorized as i) Cloud malware Injection: inject some malware application, service, or VM based machine in Cloud computing [21]. ii) Metadata spoofing: modify metadata information [22]. Younis A. Younis et al. [23] describe other security-related challenges like SLA Monitoring, management and risk analysis, heterogeneity, virtualization, trust, access control, identity management (IDM), and cross-organization management.

B. Asvija et al. [24] describe hypervisor base vulnerabilities with the use of a designed framework based on CPU, memory, and I/O firework in virtualization. Some attacks justified by the proposed technique, like GPU-based side-channel attacks, I/O channel attaches and shared memory side-channel attacks in a shared pool of resources. There will be risk analysis to analyze the performance and deficiency in deployment and measuring security traits. SVM is used to gather the possible online attack over the cloud nodes on a hypervisor level by using parameters system and network-level utilization. This approach consists of detect, remediate, recover, and defend process steps and will be able to improve 90 percent in malware and DOS detection [25].

III. PROPOSED WORK

In the proposed research, in proposed technique presented a hybrid optimization and classification algorithm for dynamic resource allocation in a multi-cloud paradigm. This hybrid algorithm is a combination of (shuffled leaf frog algorithm) SLFA with (Ubiquitous Binary Search) UBS and pure genetics theories. SLFA analyzes all the available resources from shared pool in the multi-cloud and divides them into small groups. Resources from these small groups are selected on need-based and demand-based. Need-based resources are used to complete the initial requirement of each virtual instance. Demand-based resources are required to complete a specific task on a particular VM instance, where needed a full optimization with cost efficiency. Therefore, can be borrowed the resources to meet the runtime need from VM in the same pool or from shared pool. These borrowed resources will be returned back after completion of the task. The proposed technique ensured the optimal utilization of demand-based resource with minimal cost.

A small group of resources created on factors like efficiency, throughput and cost in terms of energy-efficient mode. However, to find the best and a reliable resource pool in the defined mechanism depends on two different search methods internal (need-based) and external (demand base) search. UBS algorithm is used to achieve the best performance and minimum cost to search from everywhere with efficacy and availability of required resource pool for a specific VM in a multi-cloud environment. The flow chart in "Fig. 1" presented the detailed relationship of SLFA and UBS algorithm to attain optimization.

The particular VM generates resources demand request from a shared pool based on the following equation.

$$VRp(x) = \{vRp1, vRp2, vRp3, \dots, vRpn\}$$

Where $VRp(x)$ presents the number of requests from any VM and x may be any real number from 1 to n . The request

generated in a real time-based scenario, depending on the UBS search algorithm. The diagram "Fig. 1" elaborates the relationship between optimization and classification of the proposed dynamic resource allocation technique.

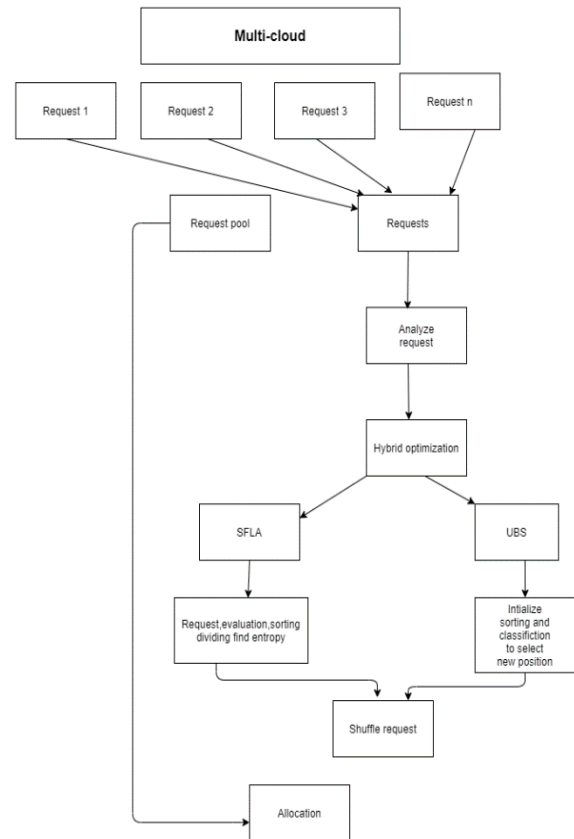


Fig. 1. UBS and SLFA Optimization Framework Flow Chart.

IV. HYBRID OPTIMIZATION AND CLASSIFICATION

Hybrid optimization depends on the deployment of SLFA and UBS in multi-cloud. This combination provides optimal dynamic resource allocation from a shared pool [26]. The hybrid deployment provides the best-optimized solution as compare to prior scheduling and scaling techniques. In this section, this technique presented a deployment of SLFA, UBS, GA, and hybrid SLFA.

A. Shuffled Leapfrog Algorithm

In the formation of subgroups, must know about the number of groups and the number of resources in each group. If the number of groups and the number of each type of resources in the group can be considered as one, the total resources will be $T = G \cdot R$. The number of best utilization of resources depends on the cost function of each group with optimal value. The whole resources divided into small categories; e.g., each type of resource must be in a group to fulfill need-based allocation.

In the division, choose the first member from the first category and second member from the latter category of the resource until the poll completes and add into G . After completion of the division, external search of a particular resource pool will start [27].

B. External Research

The metadata exploring process of the resources divided into two main stages external and internal. This step involves selecting the best optimal pair of resources to complete the initiated request involve the following steps.

Step 1: Initialization: select G and R, where G represents a small group of resources and R the number of resources. So the total number of groups with complete resources will be.

$$T = G * R$$

Step 2: Reproduction of Virtual groups: For the available resources, the sample of S virtual resources will be $v(1), v(2) \dots V(s)$.

$V(i) = \{v_{i1}, v_{i2}, \dots, v_{id}\}$ where d present the matter of decision and selection of particular resource in the group.

Step 3: Section and sorting: Sort the resources in descending order in a particular group. Then complete group will be $v(i), s(i)$ and $i = 1 \dots$ to s. the best place of the resources is $(v = Px(i))$.

Step 4: Division of resources into groups: Divide array X into Y that each of them obtains N resources.

Step 5: Resource evaluation in each group: Each Y_k where $k = 1, 2, 3 \dots G$ in each group assessed by the internal search given bellow.

Step 6: Combination of groups: After the evaluation, each group contains $(Y_1 \dots Y_G)$ the number of each type of resource for a specific pool.

Step 7: If the convergence conditions meet, then stop else; go to the fourth step of external research.

C. Internal Research

In the 5th step of the external search, the selection of group performed N time independently. After the completion of the search mechanism, the algorithm will return to global research to complete the step. Internal research steps are as follows.

Step 1: Set iG and iN to zero. Where iG counts the number of groups and iN counts the progression steps.

Step 2: $1 + iG = iG$

Step 3: $1 + iN = iN$

Step 4: Creation of subgroups: The creation of subgroups related to higher value associated with best one and lower value associated with a lower one. Value is assigned with the help of triangular probability distribution.

Step 5: Correction of the worst position: This is calculated by combining the lowest associated probability and selection leap parameter if the resource e.g., Ram, has the lowest capacity than the needed one. Now, if reassigned the value to a particular resource, then in the best case, go for step 8 of the external search; otherwise, go to step 6 of the external searches.

Step 6: calculate the resource size with the best-assigned value. If results obtained from step 5 are not better, then the size of the particular resource can be calculated as efficiency.

After reassigning value to a particular resource, if its efficiency is better than the previous one, it can be replaced with the previous one's value. Otherwise, go to the next step of internal research.

Step 7: If the new value of the resources is not according to the given need, then randomly generate a new virtual resource and replace that particular resource with a randomly generated new virtual resource in the pool.

Step 8: Update the group: After changing the value of the lowest resource with the newly virtually generated resource, sort them in descending order with the UBS algorithm's help.

Step 9: if $N > iN$ perform step three of the internal search again.

Step 10: if $G > iG$ goes to step 1 of internal search, otherwise return to the external search to combine the groups.

D. Ubiquitous Binary Search

UBS is all aware of the binary search algorithm, very complex to resolve the required query. Binary search algorithm always follows a sequential search pattern, so it is very costly and complex to get output. The cost of the algorithm is highly concerning the complexity of the search scenario. UBS can search everywhere, anytime, with any sequence from any medium or device. Hence, it is very helpful to implement in current distributed nature infrastructure like multi-cloud. In our proposed model, UBS used with no loops and no equal check. Like can find the required output by using \leq and \geq , which is very helpful to reduce review and loops and increase the efficiency of the algorithm with low cost. The low and high method or use to select with an appropriate resource for the best selection in low will always less than high and a mid to calculate the average. The cost of the algorithm will be $\lg(n)$.

```
def ubiquitous_binary_search(a,key) # a is the array and key is
the value we want to search

lo= 0
hi = a.length-1

while(hi-lo>1)
  mid = lo + (hi-lo)/2

  if a[mid]<=key
    lo=mid
  else
    hi=mid
  end
end

if (a[lo]== key)
  return lo
else
  return "value not found"
end
end
```

E. Genetic Algorithms

Genetic algorithms are used in natural selection techniques to sort out complex problems. These algorithms define a given problem as a string in the workspace and find the result with core genetic parameters in that give model.

GA algorithms are probabilistic and use natural evaluation in the propagation of results. The population of biological metrics like chromosomes used as an idle parameter to emulate GA algorithms. As the different types of resources in a resource pool, select and the best match one to overcome demand-based and need base requirements of a particular system in a multi-cloud architecture. Nature resources parameters are used to emulate and propagate the best results in the selection pool.

F. Hybrid SFLA and UBS

In the hybrid proposed solution, SFLA used with UBS qualities together and founded better results. The proposed hybrid algorithms showed more throughput, efficiency, and

cost savings in terms of execution time and turnaround time. UBS used as a searching algorithm to find the best resource for selection and get a better combination of need-based and demand-based requirements. The proposed model helpful to choose the best appropriate resource to fit in the required need of a specific VM in multi-cloud infrastructure deployment with hybrid vendor-based architecture.

The proposed model described in the form of flow chart-based implementation in “Fig. 2”.

The cost of this hybrid model in term of execution time and turnaround time calculated as

- Worst case time: $O(\lg(n))$.
- Average case time: $O(1)$.
- Average case time: $O(\lg(n))$.

Therefore, gain efficacy and energy-saving model by using the proposed hybrid algorithm for dynamic resource allocation in multi-cloud.

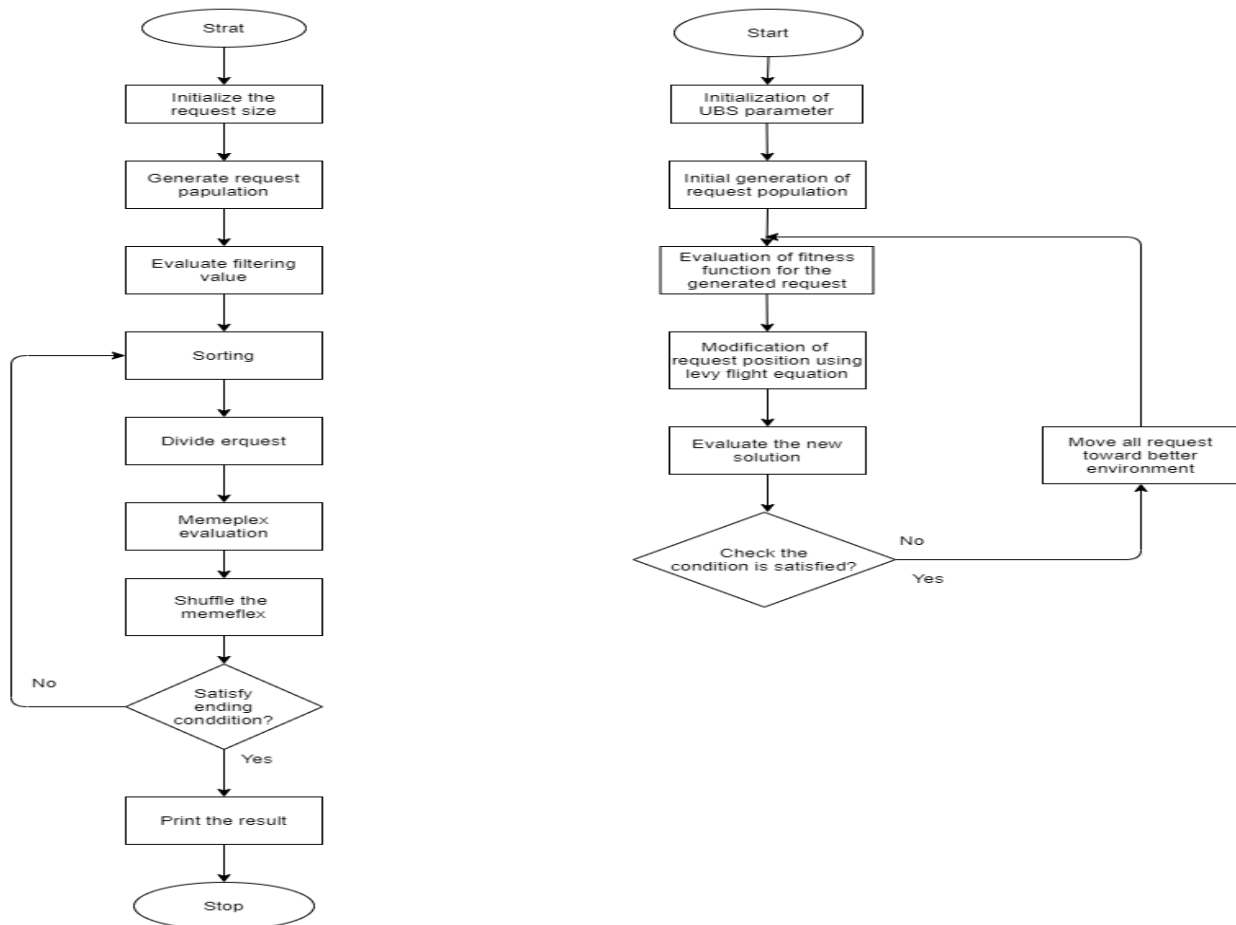


Fig. 2. Hybrid SFLA and UBS Proposed Model Flow Chart.

V. RESULTS AND DISCUSSION

In the proposed model throughput, Turnaround time, and execution time as metrics to evaluate the QoS and cost of the design algorithm.

A. Throughput

Throughput is maximum output at a certain time for a specific hardware type. It is performance measure of basic resources such as RAM, Hard drive, and CPU in a shared resources pool [28]. The throughput calculated as

$$T_t = I_t/t$$

According to the statistics shown below diagram, T_t presents the maximum output of a certain resource at a specific time. I_t is the peak value of resource under time unit "t." Results shown that the proposed hybrid algorithm present maximum throughput, which helps to create and maintain a complete resource group according to the given need. The throughput values taken at the point where maximum number of task executed by each algorithm individual. In designed case, the maximum number of tasks up to 300. The graphical presentation of different algorithms with respect to times per second is illustrated in "Fig. 3". Where "Table I" presented the numeric values of throughput in terms of per second.

B. Turnaround Time

Turnaround time presents the maximum amount of time from submitting a specific task, and its output returns to the user. It totally depends upon the function used by the developer in the code. Hence, UBS performed with less turnaround and execution time due to the absence of loop complexity in generic binary algorithms [29]. It can be calculated as

$$T(t)_{avg} = C(t) - A(t)$$

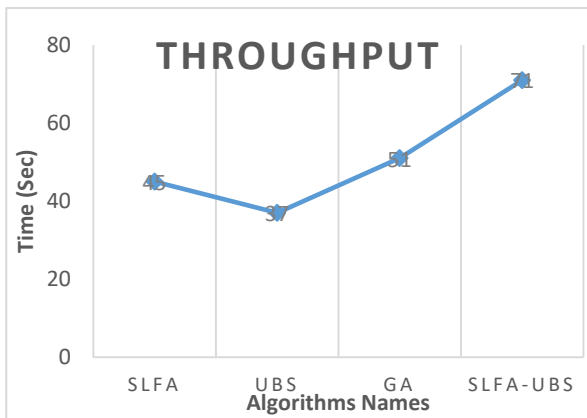


Fig. 3. Throughput Graph.

TABLE I. THROUGHPUT ON A CERTAIN DEFINE POINT WITH MAXIMUM NUMBER OF TASKS 300

Name	Time (Sec)
SLFA	45
UBS	37
GA	51
SLFA-UBS	71

Where $T(t)_{avg}$ presented average turnaround time with CT, total execution time and $A(t)$ time of arrival of a specific task. "Table II" described the turnaround time as per number of tasks for each algorithm.

Results showed that the hybrid SLFA-UBS algorithm performs continually constant with the increase of the number of tasks on the horizontal axis. The proposed model is much efficient to resolve maximum number of tasks in less turnaround time as shown in the "Fig. 4".

C. Execution Time

Total time requires executing a certain task as per user's demand. The execution time of SLFA-UBS is less, as compare to individual performances of every algorithm shown in "Table III" and "fig 5" respectively. Execution time can be calculated as

$$E(T) = E(t) - F(t)$$

Where $E(T)$ presents the computational time required to execute a specific task, $E(t)$ ending of the task, and $F(t)$ is the beginning of a specific task by the user. With respect to the number of tasks over total time to execute, SLFA-UBS performs better than any individual algorithm.

TABLE II. TURN AROUND TIME COMPARISON

Number of tasks	SLFA	UBS	GA	SLFA-UBS
50	21	25	35	13
100	23	27	37	13
150	24	28	39	15
200	25	30	40	15
250	26	31	41	17
300	27	32	42	17

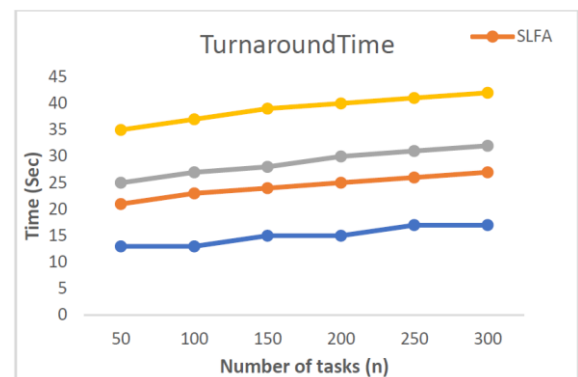


Fig. 4. Turnaround Time Graph.

TABLE III. EXECUTION TIME COMPARISON

Number of tasks	SLFA	UBS	GA	SLFA-UBS
50	19	17	15	7
100	21	23	25	12
150	27	25	23	14
200	31	32	35	15
250	34	35	38	18
300	37	39	44	20

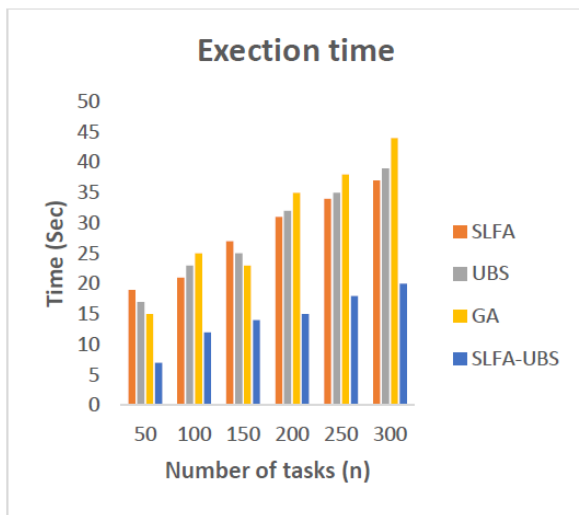


Fig. 5. Execution Time Graph.

VI. CONCLUSION

In the multi-cloud deployment dynamic resource allocation and best resource selection for particular virtual machine to fulfill the need-based resources requirement is a big challenge, and creates an adequate concern for cloud services providers. Resource optimization with minimal computational cost and throughput is also a rising concern with respect to better resource allocation in multi-cloud environment. SLFA-UBS proposed algorithm used to resolve the problems related to optimization and cost management in the selection of particular need-based resources from a shared pool in a multi-cloud paradigm. Execution of the proposed model used “CloudSim” to simulate and extract the results. The results showed that the throughput of the proposed model is 60% better than a single individual algorithm. The turnaround time and execution time had very few seconds variations near to constant as per increase in the number of tasks.

The proposed work ensured resources optimization with the dynamic provisioning techniques in the multi-cloud environment. Cost efficiency and quality of services also obtained in the summarization of this research work. Hence, we concluded that our proposed SLFA-UBS algorithm obtained better performance in optimal dynamic resource provisioning with QoS and low cost. In the future we extend our work to real time deployment of designed algorithm on cloud service providers end to enhance the archived results in test environment.

REFERENCES

- [1] Kumar, R., & Goyal, R. (2019). On cloud security requirements, threats, vulnerabilities and countermeasures: A survey. *Computer Science Review*, 33, 1-48.
- [2] Diouani, S., & Medromi, H. (2018). Green cloud computing: Efficient energy-aware and dynamic resources management in data centers. *International Journal of Advanced Computer Science and Applications*, 9(7), 124-127.
- [3] Ahmad, I., & Chang, K. (2020). Mission-critical user priority-based cooperative resource allocation schemes for multi-layer next-generation public safety networks. *Physical Communication*, 38, 100926.
- [4] Suresh, A., & Varatharajan, R. (2019). Competent resource provisioning and distribution techniques for cloud computing environment. *Cluster Computing*, 22(5), 11039-11046.

- [5] Yuan, X., Min, G., Yang, L. T., Ding, Y., & Fang, Q. (2017). A game theory-based dynamic resource allocation strategy in geo-distributed datacenter clouds. *Future Generation Computer Systems*, 76, 63-72
- [6] Morcos, M., Chahed, T., Chen, L., Elias, J., & Martignon, F. (2018). A two-level auction for resource allocation in multi-tenant C-RAN. *Computer Networks*, 135, 240-252.
- [7] Mireslami, S., Rakai, L., Wang, M., & Far, B. H. (2019). Dynamic cloud resource allocation considering demand uncertainty. *IEEE Transactions on Cloud Computing*.
- [8] Devarasetty, P., & Reddy, S. (2019). Genetic algorithm for quality of service based resource allocation in cloud computing. *Evolutionary Intelligence*, 1-7.
- [9] Haji, L. M., Zeebaree, S. R., Ahmed, O. M., Sallow, A. B., Jacksi, K., & Zeabri, R. R. (2020). Dynamic resource allocation for distributed systems and cloud computing. *TEST Eng. Manag.*, 83, 22417-22426.
- [10] Praveenchandar, J., & Tamilarasi, A. (2020). Dynamic resource allocation with optimized task scheduling and improved power management in cloud computing. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.
- [11] Kaur, S., & Bhushan, R. (2013). Review Paper on Resource Optimization of Servers using Virtualization. *International Journal of Advance Research in Computer Science and Software Engineering*, 3, 327-332.
- [12] Ding, W., Gu, C., Luo, F., Chang, Y., Rugwiro, U., Li, X., & Wen, G. (2018). DFA-VMP: An efficient and secure virtual machine placement strategy under cloud environment. *Peer-to-Peer Networking and Applications*, 11(2), 318-333.
- [13] Mohd Hairy Mohamaddiah, Azizol Abdullah, Shamala Subramaniam, Masnida Hussin “A Survey on Resource Allocation and Monitoring in Cloud Computing”, *International Journal of Machine Learning and Computing*, Vol. 4, No. 1, February 2014.
- [14] Amato, F., Moscato, F., Moscato, V., & Colace, F. (2018). Improving security in cloud by formal modeling of IaaS resources. *Future Generation Computer Systems*, 87, 754-764.
- [15] Singh, S., Jeong, Y. S., & Park, J. H. (2016). A survey on cloud computing security: Issues, threats, and solutions. *Journal of Network and Computer Applications*, 75, 200-222.
- [16] Ramachandra, G., Iftikhar, M., & Khan, F. A. (2017). A comprehensive survey on security in cloud computing. *Procedia Computer Science*, 110, 465-472.
- [17] Javanmardi, S., Shojafar, M., Persico, V., & Pescapè, A. (2020). FPPTS: A joint fuzzy particle swarm optimization mobility-aware approach to fog task scheduling algorithm for Internet of Things devices. *Software: Practice and Experience*.
- [18] Kumar, K. D., & Umamaheswari, E. (2018). Prediction methods for effective resource provisioning in cloud computing: A survey. *Multiagent and Grid Systems*, 14(3), 283-305.
- [19] Xu, X., Mo, R., Dai, F., Lin, W., Wan, S., & Dou, W. (2019). Dynamic resource provisioning with fault tolerance for data-intensive meteorological workflows in cloud. *IEEE Transactions on Industrial Informatics*, 16(9), 6172-6181.
- [20] Xu, X., Cao, H., Geng, Q., Liu, X., Dai, F., & Wang, C. (2020). Dynamic resource provisioning for workflow scheduling under uncertainty in edge computing environment. *Concurrency and Computation: Practice and Experience*, e5674.
- [21] Jensen, M., Schwenk, J., Gruschka, N., & Iacono, L. L. (2009, September). On technical security issues in cloud computing. In *2009 IEEE International Conference on Cloud Computing* (pp. 109-116). Ieee.
- [22] Jensen, M., Gruschka, N., & Herkenhöner, R. (2009). A survey of attacks on web services. *Computer Science-Research and Development*, 24(4), 185.
- [23] Younis, Y. A., & Kifayat, K. (2013). Secure cloud computing for critical infrastructure: A survey. *Liverpool John Moores University, United Kingdom, Tech. Rep.*, 599-610.
- [24] Asvija, B., Eswari, R., & Bijoy, M. B. (2019). Security in hardware assisted virtualization for cloud computing—State of the art issues and challenges. *Computer Networks*, 151, 68-92.

- [25] Veloudis, S., Paraskakis, I., Petsos, C., Verginadis, Y., Patiniotakis, I., Gouvas, P., & Mentzas, G. (2019). Achieving security-by-design through ontology-driven attribute-based access control in cloud environments. *Future Generation Computer Systems*, 93, 373-391.
- [26] Brabra, H. (2020). *Supporting management and orchestration of cloud resources in a multi-cloud environment* (Doctoral dissertation, Institut Polytechnique de Paris; Université de Sfax (Tunisie). Faculté des Sciences économiques et de gestion).
- [27] Misener, P. (2020). *Food Insecurity and College Athletes: A Study on Food Insecurity/Hunger among Division III Athletes* (Doctoral dissertation, State University of New York at Binghamton).
- [28] Kwan, A., Wong, J., Jacobsen, H. A., & Muthusamy, V. (2019, July). Hyscale: Hybrid and network scaling of dockerized microservices in cloud data centres. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)* (pp. 80-90). IEEE.
- [29] Lin, W., Peng, G., Bian, X., Xu, S., Chang, V., & Li, Y. (2019). Scheduling algorithms for heterogeneous cloud environment: main resource load balancing algorithm and time balancing algorithm. *Journal of Grid Computing*, 17(4), 699-726.

Software Engineering Ethics Competency Gap in Undergraduate Computing Qualifications within South African Universities of Technology

Senyeki M. Marebane¹

Faculty of ICT
Tshwane University of Technology
eMalahleni, South Africa

Robert T. Hans²

Department of Computer Science
Tshwane University of Technology
Soshanguve, South Africa

Abstract—Computing graduates working as software engineers are expected to demonstrate competencies in various categories of software engineering ethics as a component of non-technical skills that complement technical skills. Therefore, university programme offerings should provide opportunities for students to develop software engineering ethical competence. This study analyses curriculum documents to determine the extent to which entry-level undergraduate computing qualifications of Universities of Technology (UoTs) in South Africa provide opportunities to empower students with software engineering ethical competence. We used summative content analysis to analyze texts within the UoT computing undergraduate qualifications related to software development as retrieved from the South African Qualifications Authority database. ATLAS.ti text analysis tool was used to classify texts according to predetermined software engineering ethics categories to determine the extent to which the qualifications under study expose students to software engineering ethics. The results show that the coverage of the various categories of software engineering ethics by UoT computing qualifications for software development is insufficient, incomplete and superficial, providing only limited opportunities for prospective software engineers to develop software engineering ethical competence. Lack of adequate inclusion of software engineering ethics by UoT qualifications in South Africa deprives prospective software engineers an opportunity to develop ethical competence required to become ethically successful software engineers. Such limited exposure by software development graduates risks the development of potentially unethical software products in the software industry.

Keywords—Software engineering ethics; software engineer; technical skills; knowledge; curriculum; professional ethics; general ethics; university of technology

I. INTRODUCTION

Software Engineers (SEs) are expected to possess technical capabilities, knowledge and skills [1] along with personal or non-technical capabilities [2] necessary to meet the demands and standards of their work in developing complex software solutions. The globalized, rapidly changing world of information and communication technology (ICT) further necessitates such an all-encompassing need [3], primarily with non-technical skills, for collaborative software development environments [4]. The need for soft skills to enable SEs to

appreciate their professional ethical responsibilities towards society and the environment when applying technical skills, while serving as exemplary ethical leaders, is imperative [5]. An ICT professional should have the competencies, which consist of knowledge (what one knows), skills (what one can perform) and disposition (what personal qualities one possesses), as depicted by the competency model in Fig. 1 developed by [6].

As software engineering ethics is a critical knowledge area in today's computing world, it is important to recognize that application of hard or technical skills in software development requires a balance with soft or behavioral skills [7], [8]. In agreement, [9]–[11] assert that the possession and use of soft skills contributes more to an individual's success or failure than technical skills or intelligence. Given this, it is expected that computing qualifications for software development equip graduates with an amalgamation of the competencies as cited above, particularly those relating to personal behaviors of software engineering ethics, to escalate the ethical success of software development graduates. This is vital as an SE without a solid ethics education is a depersonalized and a mere technical instrument [12] with the potential to be misused.

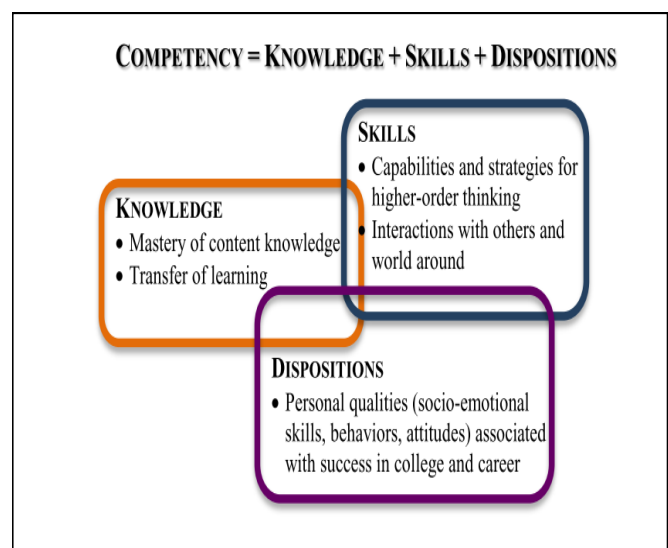


Fig. 1. Competency Model – Source [60].

However, in the recent past, based on certain reported incidents, the ethical behavior, or lack thereof, of SEs has been thrust into sharp focus. The notable unethical conduct involving SEs includes, amongst others: Volkswagen's emission scandal [13]; Uber's 'God View' app for tracking people [14]; Greyball's use of software to evade law enforcement officials [15]; Boeing's Manoeuvring Characteristics Augmentation System (MCAS) failure that led to catastrophic crashes of 737 Max 8 airplanes [16], [17]; and South Africa's Experian data breach of customer records [18]. Although popular media has reported incidents worldwide, evidence in literature to specifically provide a picture of ethical software engineering challenges in South Africa was not found by the authors.

It is evident that SEs need to increase awareness of ways that will appeal to their conscience when working in software development to ensure that the resulting software products intend good to human lives [19]. According to [20], at present, ICT professionals are not adequately empowered to competently handle ethical challenges at work, which, suggests lack of training on ethics. A worrying lack of ethics awareness as a consequence of a lack of exposure to ethics training during formal studies amongst information systems professionals in South Africa was identified by [21]. In a study on ICT graduate skill requirements, ethics, professional ethics and responsibility did not make the list of skills demanded by the South African market [22]. Furthermore, research revealed disheartening global trends of ethical breaches that have manifest in various forms of academic dishonesty and immoral behavior by computing students during their formal studies [23]–[27] despite policies and clear efforts to educate students about ethics. This suggests unethical behaviors in South African software development environments from a lack of ethical awareness, likely due to deficiencies in computing qualification curricula. While the extent of the risk stemming from a lack of ethical awareness cannot be quantified, it is likely in line with the size of software development activity in a particular country. The size of South Africa's software market in the ICT industry is briefly discussed in the next section. In South Africa, there are numerous possible causes for the dearth of software engineering ethical awareness: a lack of inclusion of ethics in the curriculum; insufficient pedagogical methods for teaching ethics; lack of ethical culture amongst academics; and industry not emphasizing ethics as a required soft skill because of an overall organizational culture devoid of ethics.

Therefore, research on computing curriculum to determine the inclusion of software engineering ethics is necessary. This paper aims to establish the opportunities of entry-level computing programme offerings by UoTs to empower prospective software development graduates with software engineering ethics competency; and subsequently reduce ethical challenges in the software development industry, advancing the development of software products that prioritize the protection of society against harm. In an effort to meet this research study's aim, the following research question will be answered: To what extent do computing qualifications for software engineering graduates offered by UoTs include ethics? The study's research objectives are as follows:

- 1) To identify software development-related entry level computing qualifications offered by UoTs registered by South African Qualifications Authority (SAQA).
- 2) To investigate if ethics learning outcomes are included in the software development curricula in the UoTs.
- 3) To determine the extent to which UoTs include ethics learning outcomes to empower graduates with ethical analysis competency.

The remainder of this paper is structured as follows: Section II presents background to the study and Section III presents a literature review. The research methodology followed by this research study is discussed in Section IV. Sections V and VI presents research results and a discussion of the results, respectively. Section VII presents the conclusion and recommendations of the study, while the limitations of this study are presented in Section VIII.

II. BACKGROUND

The South African ICT sector is the largest on the African continent, with an 8.2% contribution to the country's GDP [28]. Gartner [29] forecasted the sector spending of R303.46 billion in 2019, an increase of 3.9% from 2018, and R306,644 billion in 2020, with software spending contributing R35,850 billion. In 2019, the sector recorded the highest employment demand (close to 25%) with software development being the highest sought-after skill [30], the fastest growing, well-paying job skill in the country [31].

In South Africa, various higher learning institutions, including universities (both traditional and universities of technology [UoT]), offer computing programmes to graduate potential SEs for work in the software development industry. According to [32], of the 26 public universities in South Africa between 2016 to 2018, the six UoTs graduated a combined total of 7860 computing graduates, including those in software development, in comparison to 9105 graduates from the remaining 19 universities. The difference between these two types of universities is that UoT programme offerings typically include certificates and diplomas that are practical and career-oriented [33] [34] generally with a technological approach [35]; while traditional universities are more theory- and science-oriented [35] in their mostly degree programme offerings. Given the differences, UoTs are likely to supply the industry with a high number of SEs. For example, Tshwane University of Technology's ICT Faculty graduated a total of 2592 computing graduates between 2016 and 2018, and this made it the highest supplier of computing graduates amongst the 26 South African public universities.

As the public becomes more dependent on software driven gadgets [36], this increases the need for higher learning institutions to produce software professionals who are ethically aware and competent. In contrast, however, UoT programmes are technological, practical and career-oriented, designed to meet practical industry needs, clearly different from the theoretically-oriented programmes offered by traditional universities. This practical orientation comes with the potential exclusion of theoretical and professional issues such as ethics in the curriculum. Focus on the teaching of practical aspects of technology may lead UoTs programmes to face the same

challenge, like traditional computing courses, that fail to expose students to non-technical skills [37]. The lesser entrance requirements for UoT diploma/certificate qualifications in South Africa as compared to those of degrees at traditional universities also suggests that certain topics such as ethics may not be relevant to the curriculum at diploma/certificate level, thereby depriving students of opportunities to develop ethical awareness and the associated ethical reasoning competencies. Curriculum guidelines for computing qualifications, including those listed in Table 1, are typically designed for degree courses. Therefore, UoTs in South Africa may not necessarily recognize the need to align their entry-level qualifications with the curriculum guidelines recommended by professional bodies, which emphasize technical and professional ethics skills. UoTs cannot be exempted from producing highly ethical computing graduates because they contribute a sizeable number of SEs to the ICT sector. Based on the advocacy of the competency model in Fig. 1, one would expect that computing curricula for software development in institutions of higher learning, UoTs included, emphasize all three components of well-rounded SEs, irrespective of the type of university from which they graduate.

Analyzing curriculum documents such as syllabi and curriculum descriptors [38]–[42] specifically using content analysis [43] has become critical in curriculum research. Very few studies have been undertaken to determine the coverage of ethics by computer science programmes [44]. Furthermore, analyzing computing curriculum to determine the coverage of ethics in technology courses is important [40]. Therefore, this study is important in that it will shed light on the coverage of software engineering ethics knowledge areas by software development-related qualifications offered by South African UoTs. Its findings are likely to influence future curriculum of South African UoTs' computer science programmes. Furthermore, the outcome of the study will be of interest to software engineering and computer science curriculum development practitioners of the UoTs concerned, as the training of ethical SEs is of paramount importance for ICT employers in particular and society at large.

III. LITERATURE REVIEW

Recent years have seen a proliferation in the use of software applications and software-controlled technological devices [45]–[47] because of the benefits society derives from these technologies [2]. These require SEs to become aware and behave in a professional, social and personally responsible and ethical manner [48] given that their work has a far reach, more than even the products of other engineers [47]. Therefore, the lack of ethical awareness and ethical responsibility by computing professionals as a result of the way universities teach ethics [49], [50] challenges higher learning institutions to include and increase the coverage of ethics in their curricula [47], [51]. The inclusion of ethics in educational endeavors such as undergraduate curriculum [49] influences ethical behavior in practice [52], [53] and helps learners to develop ethical competence needed to exercise ethical autonomy [54].

Competency is an important concept in the execution of work responsibility and decision-making, especially where the work environment requires professional and technical skills, agility, dynamism and the ability to take decisions under pressure, as in the software development environments. Competency is defined by [56] as “a dynamic representation of demonstrated knowledge, understanding/ insight/ comprehension, (subject specific and generic) intellectual, practical and interpersonal skills and (ethical) values”. As SEs are confronted with decisions on competing technical, social and moral issues in the development of software, ethical competence is critical for the successful balancing of this competition. The cognitive ability which results in individual autonomy, that is, the understanding and proficient application of ethical skills at personal and organizational levels [57] in dealing with ethical problems and conflicts [55] is ethical competence. This epitomizes the actual competence required in SEs to act maturely, to responsibly apply the requisite ethical skills when confronted with ethical dilemmas, other than the humdrum regurgitation of theoretical ethical knowledge.

In recognition of the need for ethical awareness and subsequent ethical competence in software development, there has been a notable shift in computing curriculum design to recognize social, political and environmental implications of technology [58]. These changes are captured in the evolution of computing curriculum that has seen the inclusion of social, ethical and professional issues [59], observable in the pioneering CS1991 and subsequent curriculum guidelines for undergraduate degree programmes including CS2014, SE2004, SE2014 and IT2107. These curriculum guideline volumes (see Table 1) provide guidelines for the inclusion of the relevant learning outcomes on software engineering ethics by higher learning institutions. Furthermore, they clearly state that bachelor degree computing curriculum should equip students with ethical competencies [60], [61] and expose them to professional responsibilities towards society [62]. The teaching of ethics in computing has also extended to various domain specific areas such as cybersecurity, computer science and machine learning [40]. It has further translated into the inclusion of ethics as part of the professional practice knowledge area in the Software Engineering Body of Knowledge (SWEBOK) [63] which also guides the curriculum guideline volumes. In support, earlier researches, such as [48], [59], [64], [65] demonstrate various advances in promoting professional practice and ethics, including the accreditation of qualifications.

Several professional bodies – such as ACM [66] and IEEE [67] have developed SWECO to provide ethical guidelines to SEs in an effort to develop the software engineering profession, including in the teaching [49], [68], [69]. However, in South Africa, the accreditation of SEs in the field of software development is not mandatory, thereby leaving the responsibility for training future SEs on ethical matters in the hands of institutions of higher learning.

TABLE I. COMPUTING CURRICULUM GUIDELINES (ADAPTED FROM IT2017 [6])

Report	Focus of undergraduate degree programme
CS2001	Computing Curricula 2001 Computer Science
IS2002	Information Systems Model curriculum and guidelines for undergraduate degree programmes in Information Systems
SE2004	Software Engineering 2004 Curriculum for undergraduate degree programmes in Software Engineering
CE2004	Computer Engineering 2004: Curriculum guidelines for undergraduate degree programmes in Computer Engineering
IT2008	Information Technology 2008: Curriculum guidelines for undergraduate degree programmes in Information Technology
CS2008	Computer Science Curriculum 2008: Interim revision of CS 2001
IS2010	Curriculum guidelines for undergraduate degree programmes in Information Systems
SE2014	Software Engineering 2004 Curriculum for undergraduate degree programmes in Software Engineering
CE2016	Computer Engineering 2016: Curriculum guidelines for undergraduate degree programmes in Computer Engineering
IT2017	Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree programmes in Information Technology

The need for research on the inclusion of ethics in computing curriculum and how it is taught is paramount [40]. There are previous studies similar to this one that have been conducted. These studies revealed that university curricula were thin on software engineering ethics coverage [70], [71], variability on ethics topics covered [40] or ethics topics were poorly covered [72]. All these studies and several others, assessed software engineering curricula using informants' opinions [44], [73]–[75], while this study evaluated the actual curricula of each UoT in South Africa. Furthermore, the study by [72] considered ethics as a topic, separate from computer science, but rather as belonging to business studies. However, such approach contradicts the consideration that ethics is a significant knowledge area for software engineering [47], [76] and that software engineering ethics should be taught by educators that possess technical skills [36].

Given that the field of ICT evolves constantly and rapidly, the design of software engineering curricula should consider preparing SEs to be compatible with current and future technologies [6]. Considering that universities are perfectly positioned to educate students to develop the much needed ethical awareness [21], [52], this study seeks to assess the extent to which undergraduate computing programmes offered by South African UoTs give attention to software engineering ethics to allow learners to develop software engineering ethics required to competently deal with ethical dilemmas.

IV. RESEARCH METHODOLOGY

As indicated above, the objective of this paper is to establish the opportunities of entry-level computing programme offerings by South African UoTs to empower prospective SE graduates with software engineering ethics competency. That is, this study intends to determine whether or not South African UoT programmes include learning outcomes

to expose students to software engineering ethical issues, but does not determine the effectiveness of the ethics awareness. To achieve this objective, the researchers used summative content analysis, which allows keywords and/or phrases to be determined upfront and during the data analysis process. As a qualitative research methodology, content analysis is suitable for analyzing text data from a naturalistic perspective [77]. The text data to be analyzed may come from various documents, such as research articles, magazines and newspapers [78], which in this case is programme descriptor documents. Table 2 depicts the key steps to be followed in content analysis and how these were applied in this study.

The following discussion outlines the procedure that was followed in analyzing the South African UoT curricula for the inclusion of ethics learning outcomes and assessment criteria.

The researchers conducted a search on the SAQA database (<http://regqs.saq.org.za/search.php>) of selected computing undergraduate entry-level diplomas and degrees (National Qualification Framework [NQF] level 6 and 7) offered by UoTs in South Africa. SAQA is South Africa's statutory body that maintains a database of qualifications registered in line with the NQF and its sub-frameworks. The database contains qualification descriptors that specify the structure and content of each registered qualification including its purpose and learning achievements expressed as exit level outcomes and associated assessment criteria. The search criteria included Higher Education Qualification sub-framework aligned qualifications offered by the six public UoTs at NQF level 6 and 7 in the Information Technology and Computer Sciences learning subfield. The search retrieved sixteen documents that described computing qualifications offered by the six South African universities of technology: namely, Cape Peninsula University of Technology (CPUT); Central University of Technology (CUT); Durban University of Technology (DUT); Mangosuthu University of Technology (MUT); Tshwane University of Technology (TUT); and Vaal University of Technology (VUT). Of the sixteen retrieved documents, seven were excluded from consideration in this study because their specialization areas are computer networking, multimedia and computer systems engineering. The remaining nine documents describe qualifications that focus specifically on software development.

To classify the text data from the curricula, the researchers made use of categories of software engineering ethics suggested by Gotterbarn [79]. However, the researchers added an additional category entitled Structures to classify the material used as frame of reference for ethics, such as codes of ethics and codes of professional practice. The researchers considered the keywords and terms from IEEE-CS and ACM codes of ethics to search and categorize the text data from curricula documents. Some of the keywords or terms, such as reliable/reliability, best practice and professional competency, were discovered in the curriculum documents during the reading process by the researchers, justifying the use of summative content analysis in the study. The contextual use of a new keyword or phrase and its explicit description of ethics were the determinants of its inclusion in the list of keywords. Since a keyword or term may belong to more than one category, the grouping of these keywords was not restricted to

any specific category, but rather to all applicable categories, as seen from Table 3. Typically, however, the context within which a term was used in the text determined the category into which the text data would be classified.

The researchers made use of a text analysis tool, ATLAS.ti, for coding and analyzing the curricula for the inclusion of ethics learning outcomes and assessment criteria. Fig. 2 provides an example of a curriculum document coded using ATLAS.ti. Furthermore, the researchers carefully read the curriculum documents to ensure that no text was missed in the coding process and that each text was assigned to the correct category based on its contextual meaning and usage. This approach finds support from [80], who explains that qualitative researchers rely on reading the text data in the coding process.

Both authors of this article were involved in the creation of the procedure for analyzing the content of the curricula as well as the actual process of analysis. Each qualification document was coded and analyzed by each author separately. After this, the authors cross-compared their coding and analysis results to determine if there were any differences in their results. Any differences in the results were discussed to establish reasons for such, and based on the discussion a decision would be made. After the cross-validation process, a consolidated coded document and analysis results were produced for each qualification document considered. The analysis results were then used to answer the study’s research questions and draw conclusions. Fig. 3 summarizes the research approach followed in this research study, as discussed above.

TABLE II. STEPS INVOLVED IN CONDUCTING CONTENT ANALYSIS

Content analysis steps	Application of content analysis steps in this article
1. Select content to be analyzed	Documents describing qualifications that focus on software development from the South African UoTs were selected.
2. Define units and categories of analysis	The software development qualifications offered by the South African UoTs are the units of analysis of this study. The categories of analysis are given in Table 3.
3. Develop a set of rules for coding	<ol style="list-style-type: none"> 1. Only keywords/phrases that explicit describe code of ethics were considered. 2. We used the category description to categorize the search results because the keywords /phrases may not necessarily determine a category.
4. Code the text as the rules	The coding of text was done using ATLAS.ti and then verified manually to ensure that the results were placed in correct categories based on the context. We also manually went through the text to make sure that no synonyms of the keywords were missed in the search using ATLAS.ti.
5. Analyze the results and draw conclusions	The results of the search were analyzed and interpreted in order to answer this study’s research question.

TABLE III. CATEGORIES OF SOFTWARE ENGINEERING ETHICS (ADAPTED FROM [79])

Categories	Category description	Terms/Keywords
Structures	Material for frame of reference to provide ethical guidelines such as code of ethics, code of professional practice, professional standards.	Code of ethics, code of practice, code of conduct, professional practice
General Ethics	Used to regulate human interaction through obligations voluntarily accepted by an individual. The primary goal is to achieve human and society well-being that is to protect humans/society against harms.	Care, integrity, respect, privacy, avoid harm, trustworthy, fairness, honesty, security, avoid deception, accept responsibility, safety, public interest, public good, social responsibility, accuracy, property, accessibility, responsible, ethical awareness, technological impact, security concerns, ethical practice, ethical judgement, ethical approach, ethical conduct, ethicality, ethical, ethical concern, professional, professional standard, professional judgement, non-disclosure, disclosure, ethic(s), best practice, professionalism, professional competency, standard practice, negative consequences, unethical practice, unprofessional conduct, misuse, quality, reliability, reliable, technical responsibility, risk, human values, cause no harm to environment.
Professional Ethics	Prescribed by professional bodies to obligate practitioners to maintain standards of practice in line with specific level of knowledge for the benefit of the client and regulate behavior of practitioners in order to protect the profession. Specified in codes of ethics and professional practice. Similar in nature across professions, because a professional requires specialized skills to produce a product or deliver service that affect human lives, therefore use professional knowledge to cause no harm.	
Technical Ethics	Profession specific. Technical standards agreed upon in the profession to direct the acceptable performance levels in the various activities of a practice domain with an aim to cause ‘no harm’. In software engineering, they specify technical standards across the various activities of the software process. Failure to follow these standards leads to ethical issues.	

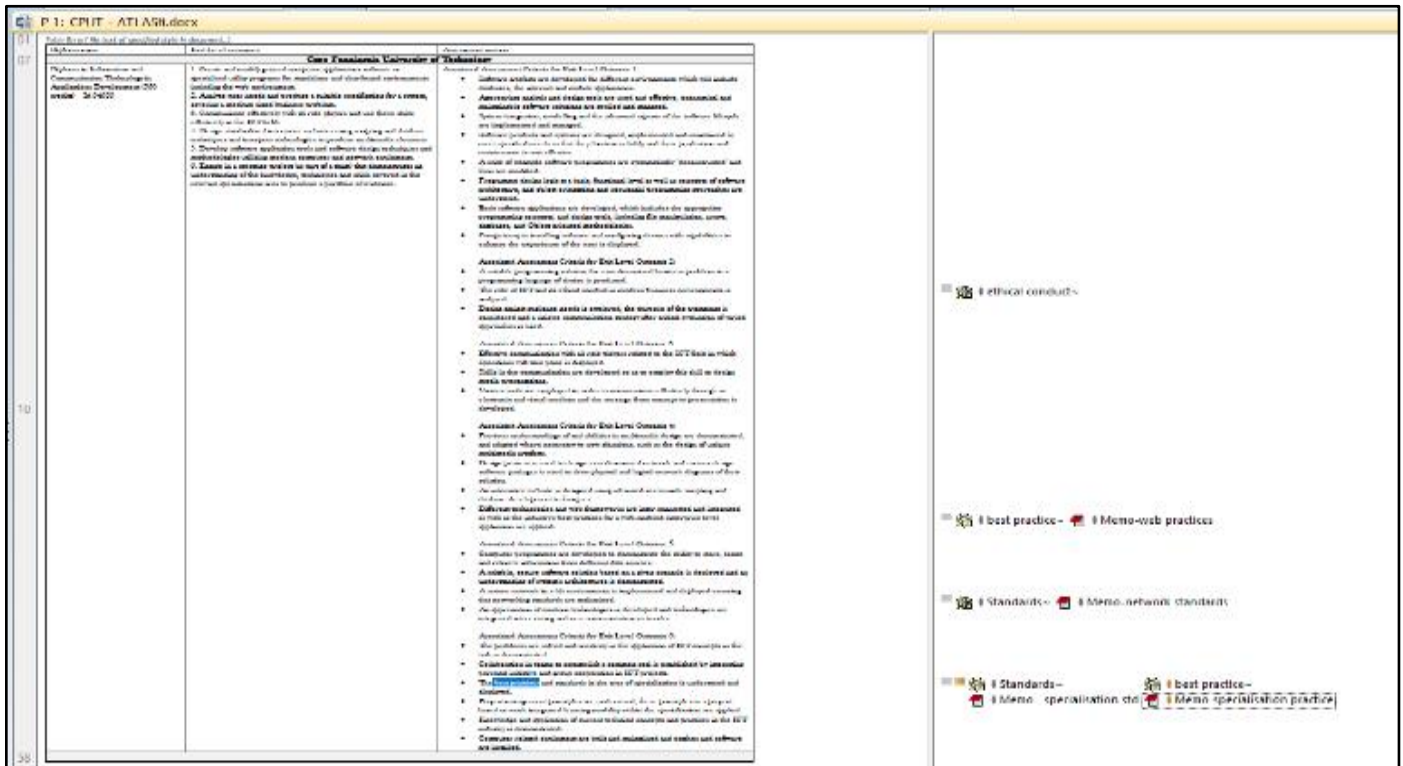


Fig. 2. An Example of a Curriculum Document that has been coded using ATLAS.ti.

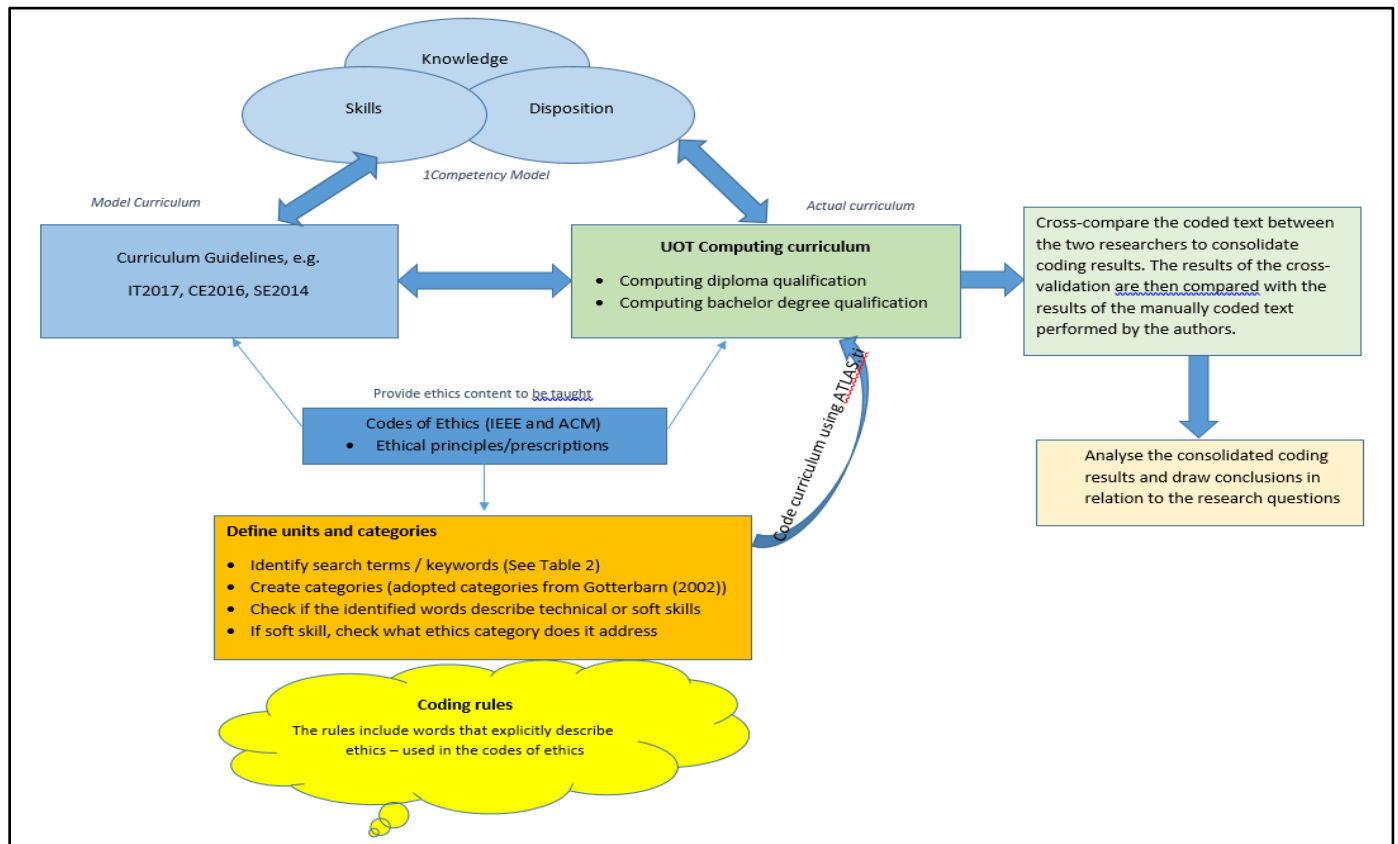


Fig. 3. This Study's Research Methodology Blueprint.

V. RESEARCH RESULTS

This section presents the results of the coding processes and analysis of the software development qualification programmes of the six South African UoTs.

A. Cape Peninsula University of Technology

The search of the document with CPUT qualifications using the keywords, which appear in Table 2 produced the results shown in Table 4. As indicated by the results in Table 3, CPUT offers one software development-related qualification, a diploma in Information and Communication Technology in Applications Development. The search did not find any match for the following categories: Structures, Professional Ethics and Values & Other. There was one match related to General Ethics. The statement mentions that students are expected to consider ethical impact of ICT in modern business environments. However, there is no indication of an ethical reference frame, such as code of ethics or code of practice. Moreover, there is no mention of how the evaluation of ICT service or product affects human lives or society. There were three matches relating to Technical Ethics. Firstly, students are expected to implement a secure network following certain networking standards, thus the standards referred to in the qualification pertain to technical standards and not to professional standards. Secondly, the references to ‘best practices’ and ‘standards’ in the learning outcomes seem to be of a technical nature considering that the statements refer to the ‘area of specialization’. Finally, the use of the term ‘best practices’ in the third match, as shown in Table 3, appears to indicate technical standards that apply to web-solutions, such as uninterrupted connection and secure data exchange. It is worth noting that the document does not indicate that human values be considered when deciding or applying technical standards.

B. Central University of Technology

The search of the curricula document of CUT using the keywords mentioned above yielded the results presented in Table 5. CUT offers one qualification related to software development, a Diploma in Information Technology. The search did not find any match for Structures, General Ethics, Technical Ethics and Values & Other categories. Two search results were categorized under Professional Ethics because the retrieved statements are more on ‘professional competencies’ of software engineers. In the first entry, as shown in Table 4, students are expected to demonstrate the level of professionalism in various ways during assessments, which include problem-solving, presentation and written examinations. However, the qualification does not clearly articulate under the learning outcomes the kind of professional ethical qualities expected from students. The second entry relates to ‘environmental sensitivity’. In practice this concept relates to ethical responsibility towards the environment (natural resources), for example trees and water, through green computing. However, in this qualification there is no mention of how students will be evaluated and how they ought to demonstrate this. Therefore, the usage of an environmental sensitivity concept in the context of this qualification seems inadequate for ethical responsibility towards environment.

TABLE IV. KEYWORDS SEARCH RESULTS OF CPUT CURRICULA DOCUMENT

Categories	Search results
Structures	No match
General Ethics	The role of ICT and its <i>ethical conduct</i> in modern business environments is analyzed.
Professional Ethics	No match
Technical Ethics	<ul style="list-style-type: none">• A secure network in a lab environment is implemented and deployed ensuring that networking <i>standards</i> are maintained.• The <i>best practices</i> and <i>standards</i> in the area of specialization is understood and displayed.• Different technologies and web frameworks are inter-connected and integrated as well as the industry's <i>best practices</i> for a web-enabled enterprise level application are applied.
Values & Other	No match

TABLE V. KEYWORDS SEARCH RESULTS OF CUT CURRICULA DOCUMENT

Categories	Search results
Structures	No match
General Ethics	No match
Professional Ethics	<ul style="list-style-type: none">• In the assessment strategy as a whole, evidence of <i>professional competencies</i> must be demonstrated through a variety of assessment methods which include case studies, problem solving assignments and strategies, portfolio of learning materials, projects and presentations, written and oral examinations, authentic practical exercises and demonstrations.• Demonstrate depth of an <i>environmentally sensitive</i>, basic business skill, and solve problems that are computer literate, numerate and be able to communicate effectively.
Technical Ethics	No match
Values & Other	No match

C. Durban University of Technology

DUT offers three software development qualifications, as shown in Table 6. The following discussion outlines the extent to which each ethics knowledge area is covered by the qualifications offered by DUT.

1) *Diploma in information and communication technology in applications development (DAD)*: This qualification does not appear to cover Structures and Technical Ethics, but does cover the other three knowledge areas, namely General Ethics, Professional Ethics and Values & Other. The two phrases – ‘ethical analysis in business situations’ and ‘ethical perspectives in business situations’ – found during the document search were categorized under General Ethics as they refer to business situations in general. Similarly, the phrase ‘ethics and social responsibility’ is aimed at guiding students to behave ethically towards society in general and hence its categorization under General Ethics. The use of the phrase ‘ethical self-awareness’ in the document refers to an individual’s internalized ethical awareness, and as a result, was classified under General Ethics. Similarly, the statement that refers to ‘ethical perspectives in business’ was classified

under General Ethics because the graduates are expected to exhibit ethical behavior in business generally. A statement relating to ‘legal’ issues of the customer/contractor relationship was classified in the Professional Ethics category because it relates to the regulation of legal issues surrounding the software development contract between the customer, which is the client company, and the software development company. However, the use of ‘legal’, especially in this context, does not necessarily relate to ethics. Lastly, the use of the terms quality and risk in the statement classified under the Values & Other category relates more to software project management than to software development, thus the classification.

2) *Diploma in information and communication technology in business analysis (DBA)*: Since the search of the keywords for this qualification returned similar results as the DAD qualification discussed in the preceding section, a similar classification process was followed. The only difference was that the search results found an entry in the DBA qualification document that relates to accounting professional practice. The statement is more aligned to accounting practice (professional

ethical conduct in accounting), which is known for enforcing and demanding ethical behavior from its professionals. It is unclear what type of ‘accounting practices’ the statement refers to – legal or ethical. However, because the statement pertains to professional ethical conduct, although in accounting, the researchers classified it in the Professional Ethics category.

3) *Bachelor of information and communication technology (BICT)*: For this qualification, the search could not find any matches for Structures, Professional Ethics, Technical Ethics or the Values & Other category. However, there were three entries that matched the General Ethics category, as shown in Table 6. The first and second entries expect graduates to determine the ‘impact of technology’ on humans (project teams included), society and business at large, and thus should be categorized under General Ethics. Since the last entry requires that students develop secure systems to safeguard harm against human beings, society and business at large, it was then categorized under the General Ethics category.

TABLE VI. KEYWORD SEARCH RESULTS OF DUT CURRICULA DOCUMENT

Diploma in Information and Communication Technology in Applications Development (DAD) (360 credits) ID: 94697	
Categories	Search results
Structures	No match
General Ethics	<ul style="list-style-type: none"> • Ethical analysis in business situations is conducted. • Different ethical perspectives in business situations are presented. • Ethics and social responsibility are demonstrated through case studies and projects. • Ethical self-awareness is demonstrated. • Different ethical perspectives in business situations are presented.
Professional Ethics	The legal aspects of customer and contractor relationship in projects are analyzed.
Technical Ethics	No match
Values & Other	Projects are managed in terms of scope, time, cost, quality , human resource, communications and risk .
Diploma in Information and Communication Technology in Business Analysis (DBA) (360 credits) ID: 97709	
Categories	Search results
Structures	No match
General Ethics	<ul style="list-style-type: none"> • Ethical analysis in business situations is conducted. • Different ethical perspectives in business situations are presented. • Ethics and social responsibility are demonstrated through case studies and projects. • Ethical self-awareness is demonstrated. • Different ethical perspectives in business situations are presented.
Professional Ethics	<ul style="list-style-type: none"> • Questionable accounting practices are identified, judgments are made, and a set of restated financials that are free of accounting concerns are produced. • The legal aspects of customer and contractor relationship in projects are analysed.
Technical Ethics	No match
Values & Other	Projects in terms of scope, time, cost, quality , human resource, communications and risk are managed.
Bachelor of Information and Communications Technology (BICT) ID: 104534	
Categories	Search results
Structures	No match
General Ethics	<ul style="list-style-type: none"> • Assess the impact of technology on individuals, organisations and society, including ethical, legal and policy issues. • Assess the impact of technology on group work in project teams, collaborating in IT-related projects. • Implement the security features and the various levels of security in existing applications.
Professional Ethics	No match
Technical Ethics	No match
Values & Other	No match

D. Mangosuthu University of Technology

There were no matches found in the document of keywords or phrases that describe software engineering ethics in the qualification for MUT. This indicates that the qualification is devoid of software engineering ethics coverage. The lack of coverage of ethics knowledge areas by this qualification means MUT produces software engineers who lack software engineering ethics.

E. Tshwane University of Technology

TUT offers two software development-related qualifications: a Diploma in Computer Science and a Diploma in Informatics. The search for the keywords given in Table 2 yielded no results for the Informatics Diploma qualification, indicating that the qualification does not cover software engineering ethics.

1) *Diploma in computer science*: Table 7 shows no entries found for two categories, Structures and Values & Other. The statements relating to ‘security concerns’ pertain to security issues in software solutions to deal with potential security threats to users in general; hence their classifications under General Ethics. The search results, which relate to software solution ‘standard practices’ were classified under Professional Ethics. The statement, which requires graduates to include ‘quality’ related factors in their software development process was classified under Technical Ethics as necessitated by the description of the category in Table 2.

F. Vaal University of Technology

Table 8 shows the outcome of the search of the document related to the Information Technology Diploma qualification offered by VUT. The search did not yield any entries related to Structures, General Ethics, Professional Ethics and Values & Other categories. All four search results were classified under Technical Ethics because they are about technical issues such as technical skills and technical operations. However, the use of the term ‘technical skill’ here is too broad and is about the implementation of an ‘effective’ solution, which may not necessarily be an ethical one, or the professional might not have adhered to ethical behavior in the implementation.

TABLE VII. KEYWORDS SEARCH RESULTS OF TUT CURRICULA DOCUMENT

Diploma in Computer Science (360 credits) SAQA ID: 109017	
Categories	Search results
Structures	No match
General Ethics	<ul style="list-style-type: none"> Discuss <u>security concerns</u> in Web and mobile applications. Explain proposed solutions to mitigate <u>security concerns</u>.
Professional Ethics	Propose and design database solutions using <u>standard practices</u> .
Technical Ethics	Incorporate <u>quality</u> enhancement factors in the programming process.
Values & Other	No match

TABLE VIII. KEYWORDS SEARCH RESULTS OF VUT CURRICULA DOCUMENT

Diploma in Information Technology (360) SAQA ID: 101144	
Categories	Search results
Structures	No match
General Ethics	No match
Professional Ethics	No match
Technical Ethics	<ul style="list-style-type: none"> Utilise the required <u>technical skills</u> to effectively implement the designed solutions in a distributed Information Technology (IT) environment. Utilise the required <u>technical skill</u> to design and implement solutions in data communications, networks and the internet environment. Use hardware to its full potential by understanding the <u>technical</u> operation of hardware and to control it on a low level. Utilise business skills effectively in the application of <u>technical skills</u> in an IT business environment.
Values & Other	No match

VI. DISCUSSION OF RESEARCH RESULTS

This section discusses the research results to establish the inclusion of software engineering ethics by individual qualifications and the level of by the qualifications combined, which were presented in the previous section. A summary on the level of coverage of each ethics knowledge area is also presented.

A. Cape Peninsula University of Technology

The qualification offered by CPUPT seems to be producing graduates who are more technically inclined but lean in other ethical knowledge areas, such as general ethics, professional ethics and ethical values. A lack of reference to structures of ethical aspects in curriculum deprives students of opportunities to learn from an organized and focused source of ethical guidelines.

B. Central University of Technology

This qualification only emphasizes professional ethics-related matters while ignoring other knowledge areas on ethics. These findings suggest that the qualification neglects to develop students with balanced knowledge in ethical issues. Even though the qualification seems to be focused on professional ethics, its focus area is still incomplete in that the expected professional ethical qualities are not clearly outlined and the qualification does not indicate how environmental sensitivity issues will be determined.

C. Durban University of Technology

1) *Diploma in information and communication technology in applications development (DAD)*: Even though this qualification does not cover all the knowledge areas of ethics, its coverage is still much fuller than the qualifications offered by CPUPT and CUT in that it covers three of the five ethics knowledge areas. Secondly, the concept of ethics is clearly specified, especially in specific ethical concepts such as self-awareness, social responsibility, ethical perspectives and analysis, clarifying aspects of ethics to which students will be exposed. However, it does not specify the structures from

which ethical knowledge and awareness will be derived. This potentially weakens the efforts of this qualification to instill ethics knowledge in students.

2) *Diploma in information and communication technology in business analysis (DBA)*: Since the search results are mirror images of the DAD qualification, this qualification's coverage of ethical issues is also better than the qualifications offered by CPUT and CUT.

3) *Bachelor of information and communication technology (BICT)*: Unlike the other two qualifications offered by DUT, this qualification falls short in the coverage of ethical issues in four categories, as mentioned above. Although the qualification clearly states that impact of technology on individuals, organizations and society will be assessed, which is a plausible start in terms of high level intention of ethical and legal considerations, catered to better than in CPUT and CUT qualifications, it fails to identify the specific areas of ethics that relate to software engineering, thereby obscuring the direct exposure of students to the specific categories of ethics that can develop ethical competence and reasoning.

D. Tshwane University of Technology

1) *Diploma in computer science*: Although this qualification covers three of the five knowledge areas of ethics, its coverage is minimal. It neglects to include any ethical frame of reference, it lacks focus on the topics covered in ethics, and the identified ethics areas are narrow. Firstly, the non-inclusion of an ethics frame of reference, that is structures, exposes a gap in terms of guidelines that could potentially develop students' ethical competencies. Secondly, the ethics categories identified in the qualification focus only on three select items of the development of software technology: secure web/mobile applications; standards in designing databases; and quality enhancement in programming. Since this is applicable only on specific technical applications, it cannot be regarded as sufficiently full coverage of the three categories of software engineering ethics, let alone the rest of the categories. This then casts doubt on the ability of the inclusion of ethics in this qualification to develop ethical competence. If it does, it will only pertain to the three specified technologies, leaving students ethically incompetent on general activities of the software process.

2) *Diploma in informatics*: This qualification has no coverage of ethics matters. This is concerning because the qualification trains business and systems analysts who lack ethical awareness in these important ethical knowledge areas. Such a lack of ethical awareness may result in unethical conduct by programme graduates.

E. Vaal University of Technology

Generally, as all technical qualifications intend to impart technical knowledge, the mentioning of technical skills in this qualification does not necessarily refer to specific categories of ethics. The search results show that this qualification does not include any category of ethics to be learnt by students.

Therefore, as students graduating from this qualification are at risk of graduating with only technical knowledge, their lack of ethical competence may jeopardize their contribution to the ethical success of software projects in which they work.

F. Level of Coverage of Knowledge areas by South African UoTs

Fig. 4 summarizes the level of coverage of ethics knowledge areas by South African UoTs qualifications. The following discussion summarizes the research results with regard to level of coverage of each ethics knowledge area.

- Structures in the form of a code of ethics or code of practice that provides ethical guidelines are not covered by any qualifications of the UoTs. This suggests that UoTs graduate software engineers who lack an organized and professional frame of reference about ethical issues. Teaching software engineering ethics requires reference to structures; therefore, it is worrying that the UoTs do not teach students about the various categories of software engineering ethics.
- General Ethics is covered by 55% of the qualifications (i.e., five qualifications cover this category). This means that the remainder, about 45% of the qualifications are failing to cover this knowledge area. The coverage of general ethics by these qualifications is narrow; it is difficult to locate them within the major activities of the software process. In as much as this category is covered, it does not bring the necessary value to the qualification in terms of providing sufficient opportunities for students to learn the ethics aligned with the software process and to be justified as 'software engineering ethics'. The coverage of general ethics in the qualifications does not appear to lead to the advancement of human well-being as students are not taught personal responsibility of avoiding harm to humanity.
- Professional Ethics is covered by just over 40% of the UoT qualifications (only four qualifications cover this knowledge area), while a similar study by [40] found that this topic was covered by mere 22% of the qualification investigated. Our finding implies that the majority of South African UoT software development related qualifications do not train software engineers on professional ethical behavior. However, even some of these qualifications that cover this knowledge area (for example, DUT-DAD, DUT-DBA and TUT-DCS) do not cover it adequately or are ambiguous in their coverage, as discussed in subsections IV (c) (i), (c) (ii) and (e) (i). The lack of adequate coverage of professional ethics by South African UoTs may result in software engineers who are unable to fully understand their professional ethical responsibility towards clients, employers and society.
- Technical Ethics is covered by 33% (three) of the nine qualifications. This means a majority of the UoT qualifications in this regard produce professionals who are more likely to neglect to follow technical standards intended to guide them to avoid harm to individuals

and society. In as much as some of the qualifications conservatively refer to ‘standards of practice’, they rarely refer to this or fail to cite examples of such standards.

- Values are covered by 22% of the qualifications. This means that almost 80% of the qualifications neglect to cover this knowledge area, graduating software professionals who lack a moral compass to guide them in technical ethical decisions.

Only three of the nine qualifications, namely DUT-DAD, DUT-DBA and TUT-DCS, cover three of the five ethics knowledge areas. CPUT’s qualification covers two knowledge areas, while the two qualifications offered by CUT and VUT cover only one knowledge area each. The qualification offered by MUT and the DIM qualification offered by TUT do not address ethical issues whatsoever.

This study set out to meet the following research objectives:

1) To identify software development-related entry level computing qualifications offered by UoTs which are registered by SAQA. This objective was met in Section IV, where nine entry-level computing qualifications were identified and retrieved from the SAQA website.

2) To investigate if ethics learning outcomes are included in the software development curricula in the UoTs. This objective was met in Sections V and VI as UoT qualifications were searched for coverage of ethics knowledge areas and results were presented and discussed in the preceding subsections of Section VI.

3) To determine the extent to which UoTs include ethics learning outcomes to empower graduates with ethical competency. The discussion in this section outlined the extent to which UoTs include software engineering ethics in their software development qualifications. The discussion also provided an answer to this study’s research question.

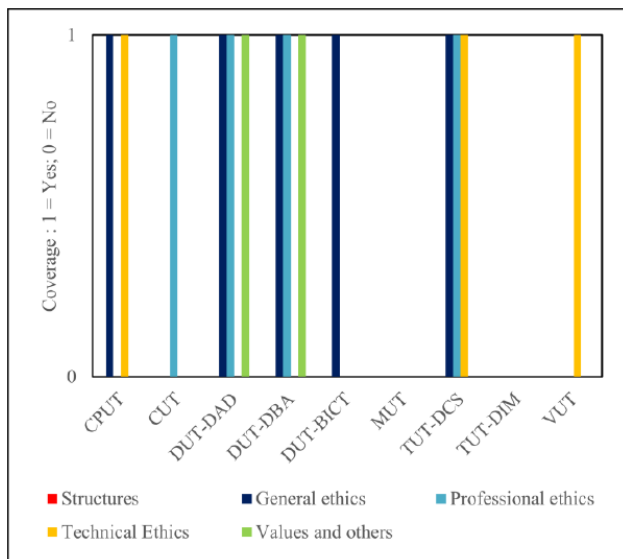


Fig. 4. Coverage of Ethics Knowledge Areas by South African UoTs.

VII. CONCLUSION AND RECOMMENDATIONS

The discussion in the preceding section points out that computing qualifications for software development offered by South African UoTs *do not* sufficiently cover the various categories of software engineering ethics. All the reviewed qualifications fell short of full coverage of the five categories of software engineering ethics and demonstrated only a superficial approach on the categories covered. Disturbingly, no frames of reference such as codes of ethics are used to guide their teaching of ethics. It is also not clear if the ethics learning outcomes are taught within a standalone module or are spread across the various modules. This potentially deprives students of opportunities for learning software engineering ethics in alignment to the various activities of the software process. Furthermore, this study’s findings concur with the assertion of [47] that the attention given to software engineering ethics by curricula is inadequate. The suggestion by [50] that computing professionals are inclined not to assume ethical responsibility because of the way we teach software engineering ethics is in agreement with the findings of this study. Educational opportunities are well situated to assist prospective SEs in developing ethical competencies for ethical software development. The inclusion of ethics learning outcomes in software engineering curricula can: empower prospective SEs to act ethically, improve the software engineering profession and lead to the development of ethical software products. This is true, especially if the learning outcomes are organized to include (a) general, technical and professional ethics and values knowledge;(b) skills required to apply the knowledge; and c) personal behaviors that bring about excellence in applying skills and knowledge pertaining to software engineering ethics, in accordance with the applicable competency frameworks such as in Fig. 1.

Moreover, the picture painted by these findings on the coverage of ethics knowledge areas by South African UoTs has some implications for the software industry and society and is somewhat concerning because UoTs are the biggest suppliers of SEs in South Africa. The graduation of SEs with a lack of knowledge and skills in various areas of ethics deprives employers of multi-skilled employees, and warrant a concern for the software industry and the society. The unethical incidents mentioned in Section I of the study might not have hit South African shores at large scale yet, but the South African situation may soon escalate if worrying gaps in UoT curricula are not addressed as soon as possible. The inability of universities to produce graduates with suitable skills as identified by [22] and also confirmed by the results, continues to deprive society and the software industry in particular, of much-needed multi-skilled and ethical SEs.

Echoing the calls of [47] and [51], this study implores South African UoTs to increase the coverage of ethics in their curricula, especially within the identified categories of Structures, Professional Ethics, Technical Ethics and Values. Such coverage should be aligned to the knowledge areas that are required for the spectrum of activities in a software process, including software engineering management. This will not only ensure that curricula remain in step with competency models in terms of ethics coverage but will also produce graduates with balanced ethical knowledge for the benefit of society and

certainly the profession. Adoption of competency models can bring an organized approach to the inclusion of ethics learning outcomes in the curricula to facilitate the acquisition and development of such competencies by South African UoT students. The time has come for South African higher education, the South African software industry and South African professional bodies to synergize efforts to align software development curricula to professional bodies guidelines, accrediting courses for software engineers to bring alignment with other professional engineering disciplines as alluded to by [81][81]. Qualifications accrediting bodies in South Africa, such as the Council on Higher Education, can contribute significantly by ensuring that computing qualifications for software development include software engineering ethics competency.

VIII. LIMITATIONS AND FUTURE STUDIES

Content analysis by its definition involves subjective interpretation [77] of the text under consideration and therefore possesses and inherent limitation. Furthermore, as there is substantive overlap between the ethics categories, the categorization of text was a subjective issue. Despite these limitations, however, this study has shown a spotlight on the insufficient coverage of ethics knowledge areas within the qualifications offered by South African UoTs. To address the abovementioned limitations, a future study could analyze study materials used by South African UoTs for the abovementioned qualifications to assess the level of coverage of ethics knowledge areas addressed in the material themselves similar to an international study by [40].

REFERENCES

- [1] J. G. Rivera-Ibarra, J. Rodríguez-Jacobo, J. A. Fernández-Zepeda, and M. A. Serrano-Vargas, "Competency framework for software engineers," in Software Engineering Education Conference, Proceedings, 2010, pp. 33–40, doi: 10.1109/CSEET.2010.21.
- [2] Y. Sedelmaier and D. Landes, "Software engineering body of skills (SWEBOS)," IEEE Glob. Eng. Educ. Conf. EDUCON, no. April, pp. 395–401, 2014, doi: 10.1109/EDUCON.2014.6826125.
- [3] M. Zahedi, M. Shahin, and M. Ali Babar, "A systematic review of knowledge sharing challenges and practices in global software development," Int. J. Inf. Manage., vol. 36, no. 6, pp. 995–1019, 2016, doi: 10.1016/j.ijinfomgt.2016.06.007.
- [4] F. Ahmed, "Software requirements engineer: An empirical study about non-technical skills," J. Softw., vol. 7, no. 2, pp. 389–397, 2012, doi: 10.4304/jsw.7.2.389-397.
- [5] A. Harris and M. Lang, "Incorporating Ethics and Social Responsibility in IS Education," J. Inf. Syst. Educ., vol. 22, no. 3, pp. 183–190, 2011, [Online]. Available: <http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authType=crawler&jml=10553096&AN=69713585&h=2IqrlxqAX3mY6CMHvX7E3jmFeST2qp7NoJEez765uemfcuUJ2gixNWDj%2F2l0iDZ0h1QTbbo%2FH1%2FdGOKV2HOnCQ%3D%3D&crl=c>.
- [6] ACM/IEEE, "Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree Programs in Information Technology," 2017. [Online]. Available: <http://www.acm.org.colorado.idm.oclc.org/binaries/content/assets/education/it2017.pdf>.
- [7] T. L. Lewis, W. J. Smith, F. Bélanger, and K. V. Harrington, "Are technical and soft skills required?: The use of structural equation modeling to examine factors leading to retention in the cs major," ICER'08 - Proc. ACM Work. Int. Comput. Educ. Res., pp. 91–99, 2008, doi: 10.1145/1404520.1404530.
- [8] Y. Sedelmaier and D. Landes, "Practicing soft skills in software engineering: A project-based didactical approach," Comput. Syst. Softw. Eng. Concepts, Methodol. Tools, Appl., pp. 232–252, 2017, doi: 10.4018/978-1-5225-3923-0.ch011.
- [9] M. B. Khan and S. Kukalis, "MIS professionals: Education and performance," Inf. Manage., vol. 19, no. 4, pp. 249–255, 1990, doi: 10.1016/0378-7206(90)90034-F.
- [10] H. Coleman Jr. and M. W. Argue, "Mediating With Emotional Intelligence: When 'Iq' Just Isn't Enough.," Disput. Resolut. J., vol. 70, no. 3, pp. 15–24, 2015, [Online]. Available: <http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=111744028&site=ehost-live>.
- [11] D. Goleman, Emotional Intelligence Why it can matter more than IQ. New York: Bantam Books, 2005.
- [12] G. Gonzalo, M. R. Gonz, and A. Fraga, "Ethical Responsibility of the Software Engineer," in Philosophical Foundations on Information Systems Engineering (PhiSE) 2006, 2006, no. 1, pp. 727–736.
- [13] R. Hotten, "Volkswagen : The scandal explained," BBC News, 2015.
- [14] S. Evans and E. Zolfagharifard, "Uber executive used 'God View' to track journalist without consent, claims report | Daily Mail Online," 2014. [Online]. Available: <http://www.dailymail.co.uk/sciencetech/article-2841329/Uber-investigates-executive-spying-customer-Manager-used-God-View-tool-track-journalist-without-consent-claims-report.html>.
- [15] J. C. Wong, "Greyball: how Uber used secret software to dodge the law," The Guardian, 2017.
- [16] BBC, "Boeing admits knowing of 737 Max problem," BBC News, 2019. <https://www.bbc.com/news/business-48174797> (accessed Jun. 03, 2019).
- [17] A. Buncombe, "Boeing 737 MAX: Company reveals new software problem detected in jets which 'must be fixed before planes can fly,'" Independent News, 2019. <https://www.independent.co.uk/news/world/americas/boeing-737-max-jets-ceo-new-software-problem-a8855841.html> (accessed Jun. 03, 2019).
- [18] A. Moyo, "Experian hacked, 24m personal details of South Africans exposed | ITWeb," ITWeb, 2020. <https://www.itweb.co.za/content/rxP3jqBmNzpMA2ye> (accessed Aug. 20, 2020).
- [19] L. A. Maxwell, "/home/oscar/GDrive/DATS2MS/LDATS2840 Memoire/Literature/Techniques/Blei2001 Latent Dirichlet allocation.pdf," Educ. Week, vol. 32, no. 5, p. 1, 2012, doi: 10.1162/jmlr.2003.3.4-5.993.
- [20] Y. Al-Saggaf and O. K. Burmeister, "Improving skill development: An exploratory study comparing a philosophical and an applied ethical analysis technique," Comput. Sci. Educ., vol. 22, no. 3, pp. 237–255, 2012, doi: 10.1080/08993408.2012.721073.
- [21] M. Charlesworth and A. D. Sewry, "South African IT industry professionals' ethical awareness : an exploratory study," in Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries, 2004, pp. 269–273.
- [22] A. P. Calitz, J. H. Greyling, and M. D. M. Cullen, "South African Industry ICT Graduate Skills Requirements," South. African Comput. Lect. Assoc., no. 1, pp. 25–26, 2014.
- [23] K. Ali, R. Salleh, and M. Sabdin, "A study on the level of ethics at a Malaysian private higher learning institution: comparison between foundation and undergraduate technical-based," Int. J. Basic and, Appl. ..., vol. 10, no. 05, pp. 20–29, 2010, [Online]. Available: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Study+on+the+Level+of+Ethics+at+a+Malaysian+Private+Higher+Learning+Institution+:++Comparison+between+Foundation+and+Undergraduate+Technical-based+Students#0>.
- [24] R. G. Ledesma, "Academic Dishonesty among Undergraduate Students in a Korean University," Res. World Econ., vol. 2, no. 2, pp. 25–35, 2011, doi: 10.5430/rwe.v2n2p25.
- [25] M. Nejati, R. Jamali, and M. Nejati, "Students' ethical behavior in Iran," J. Acad. Ethics, vol. 7, no. 4, pp. 277–285, 2009, doi: 10.1007/s10805-010-9101-4.
- [26] G. P. De Bruin and H. Rudnick, "Examining the cheats: The role of conscientiousness and excitement seeking in academic dishonesty," South African J. Psychol., vol. 37, no. 1, pp. 153–164, 2007, doi: 10.1177/008124630703700111.
- [27] G. Finchilescu and A. Cooper, "Perceptions of Academic Dishonesty in a South African University: A Q-Methodology Approach," Ethics Behav., vol. 28, no. 4, pp. 284–301, 2018, doi: 10.1080/10508422.2017.1279972.
- [28] Inserta, "Sector Skills Plan 2019/20," 2019.

- [29] Gartner, "Gartner Forecasts IT Spending in South Africa Will Grow 3.9% in 2019," Newsroom Press Releases, Jul. 24, 2019.
- [30] Career Junction, "Executive Summary December 2019," 2019.
- [31] Business Tech, "This is the fastest growing job skill in South Africa right now – and it pays really well," Business Tech, Aug. 15, 2018.
- [32] DHET, "HEMIS Data Reports 2018," 2018. [Online]. Available: <https://www.dhet.gov.za/SitePages/Higher-Education-Management-Information-System.aspx>.
- [33] Y. Farham, "What is the difference between a university, a university of technology and a TVET college? [blog]," Oxford University Press. 2015, [Online]. Available: <http://blog.oxford.co.za/what-is-the-difference-between-a-university-a-university-of-technology-and-a-tvet-college/>.
- [34] Bridge, "Universities of Technology," 2015. <https://www.bridge.org.za/knowledge-hub/psam/he/universities-of-technology/> (accessed Dec. 10, 2020).
- [35] T. Matiki, "The university of technology versus the traditional universities. Is the gap being closed?," *Mediterr. J. Soc. Sci.*, vol. 5, no. 23, pp. 2126–2128, 2014, doi: 10.5901/mjss.2014.v5n23p2126.
- [36] D. Gotterbarn and K. Miller, "Computer ethics in the undergraduate curriculum: case studies and the joint software engineer's code," *J. Comput. Sci. Coll.*, vol. 20, no. 2, pp. 156–167, 2004.
- [37] J. D. Tvedt, R. Tesoriero, and K. A. Gary, "The software faculty: An undergraduate computer science curriculum," *Comput. Sci. Educ.*, vol. 12, no. 1–2, pp. 91–117, 2002, doi: 10.1076/csed.12.1.91.8213.
- [38] F. Chong, "The Pedagogy of Usability: An Analysis of Technical Communication Textbooks, Anthologies, and Course Syllabi and Descriptions," *Tech. Commun. Q.*, vol. 25, no. 1, pp. 12–28, 2016, doi: 10.1080/10572252.2016.1113073.
- [39] Y. F. Yeh, S. Erduran, and Y. S. Hsu, "Investigating Coherence About Nature of Science in Science Curriculum Documents: Taiwan as a Case Study," *Sci. Educ.*, vol. 28, no. 3–5, pp. 291–310, 2019, doi: 10.1007/s11191-019-00053-1.
- [40] C. Fiesler, N. Garrett, and N. Beard, "What dowe teach whenwe teach tech ethics? a syllabi analysis," *Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE*, pp. 289–295, 2020, doi: 10.1145/3328778.3366825.
- [41] E. Lavrenteva and L. Orland-Barak, "The treatment of culture in the foreign language curriculum: an analysis of national curriculum documents," *J. Curric. Stud.*, vol. 47, no. 5, pp. 653–684, 2015, doi: 10.1080/00220272.2015.1056233.
- [42] J. Saltz et al., "Integrating ethics within machine learning courses," *ACM Trans. Comput. Educ.*, vol. 19, no. 4, 2019, doi: 10.1145/3341164.
- [43] D. T. Stern, "Practicing what we preach? An analysis of the curriculum of values in medical education," *Am. J. Med.*, vol. 104, no. 6, pp. 569–575, 1998, doi: 10.1016/S0002-9343(98)00109-0.
- [44] C. L. Spradling, L. K. Soh, and C. J. Ansoorge, "Ethics training and decision-making: Do computer science programs need help?," *SIGCSE'08 - Proc. 39th ACM Tech. Symp. Comput. Sci. Educ.*, pp. 153–157, 2008, doi: 10.1145/1352135.1352188.
- [45] L. F. Capretz, "Personality Types in Software Engineering," *Int. J. Hum. Comput. Stud.*, vol. 58, pp. 207–214, 2003.
- [46] U. Hansmann, L. Merk, M. S. Nicklous, and T. Stober, *Pervasive Computing: The Mobile World*, 2nd ed. Springer Science & Business Media, 2003.
- [47] J. Jia and J. Xin, "Integration of ethics issues into software engineering management education," *ACM Int. Conf. Proceeding Ser.*, pp. 33–38, 2018, doi: 10.1145/3210713.3210725.
- [48] D. L. Parnas, "The Professional Responsibilities of Software Engineers," in *In IFIP Congress (2)*, 1994, pp. 332–339.
- [49] M. C. Sabin, S. Peltsverger, C. Tang, and B. M. Lunt, "ACM/IEEE-CS information technology curriculum 2017: a status update," 2016.
- [50] D. Gotterbarn, "Informatics and professional responsibility," *Sci. Eng. Ethics*, vol. 7, no. 2, pp. 221–230, 2001, doi: 10.4324/9781315259697-27.
- [51] B. Kitchenham, D. Budgen, P. Brereton, and P. Woodall, "An investigation of software engineering curricula," *J. Syst. Softw.*, vol. 74, no. 3, pp. 325–335, 2005, doi: 10.1016/j.jss.2004.03.016.
- [52] K. Munro and J. Cohen, "Ethical Behavior and Information Systems Codes : The Effects of Code Communication, Awareness, Understanding, and Enforcement," *ICIS 2004 Proc.*, 2004.
- [53] E. A. Voutsas, K. V. Siakas, K. S. Nisioti, and M. Ross, "A Survey of Informatics Students' Awareness on Ethical Issues," in *Proceedings of the 11th International Conference on Software Process Improvement - Research into Education and Training*, 2006, pp. 139–150.
- [54] I. Kavathatzopoulos, "Moral leadership in action: building and sustaining competence in European organizations," 2000.
- [55] I. Kavathatzopoulos, "The use of information and communication technology in the training for ethical competence in business," *J. Bus. Ethics*, vol. 48, no. 1, pp. 43–51, 2003, doi: 10.1023/B:BUSI.0000004366.08853.72.
- [56] R. Wagenaar, "Competences and learning outcomes: a panacea for understanding the (new) role of Higher Education?," *Tuning J. High. Educ.*, vol. 1, no. 2, pp. 279–302, 2014, doi: 10.18543/tjhe-1(2)-2014pp279-302.
- [57] M. V. Nicoli, "A Web-Based Questionnaire of Ethical Skills," 2008.
- [58] A. Harris, M. Lang, D. Yates, and S. E. Kruck, "Incorporating Ethics and Social Responsibility in IS Education.," *J. Inf. Syst. Educ.*, vol. 22, no. 3, pp. 183–190, 2011.
- [59] C. Huff and C. D. Martin, "Computing Consequences: A Framework for Teaching Ethical Computing," *Commun. ACM*, vol. 38, no. December, pp. 75–84, 1995, doi: 10.1145/219663.219687.
- [60] ACM/IEEE Computer Society Task Group on Information Technology Curricula, *Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree Programs in Information Technology A Report in the Computing Curricula Series Task Group on Information Technology Curricula*. 2017.
- [61] ACM/IEEE, "SE 2014: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering," *Computer (Long. Beach. Calif.)*, vol. 48, no. 11, pp. 106–109, 2015, doi: 10.1109/mc.2015.345.
- [62] ACM/IEEE, "Computer Science 2013: Curriculum Guidelines for Undergraduate Programs in Computer Science. CS ACM Final Report 2013," 2013.
- [63] P. Bourque and R. E. Fairley, *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0*. IEEE Computer Society Press, 2014.
- [64] F. Bott, A. Coleman, J. Eaton, and D. Rowland, *Professional issues in Software engineering*, 3rd ed. Boca Raton: CRC Press, 2001.
- [65] F. B. Aydemir and F. Dalpiaz, "A Roadmap for Ethics-Aware Software Engineering," *2018 IEEE/ACM Int. Work. Softw. Fairness*, pp. 15–21, 2018, doi: 10.23919/FAIRWARE.2018.8452915.
- [66] ACM, "ACM Code of Ethics and Professional Conduct," *Commun. ACM*, vol. 35, no. 5, pp. 94–99, 1992, doi: 10.1145/3274591.
- [67] IEEE-CS, "Code of Ethics [IEEE-CS/ACM Joint Task Force on Software Engineering Ethics and Professional Practices," 1999. [Online]. Available: <https://www.computer.org/education/code-of-ethics>.
- [68] E. Towell, "Teaching ethics in the software engineering curriculum," in *Software Engineering Education Conference, Proceedings, 2003*, vol. 2003-Janua, pp. 150–157, doi: 10.1109/CSEE.2003.1191372.
- [69] E. Towell and J. B. Thompson, "A further exploration of teaching ethics in the software engineering curriculum," in *Software Engineering Education Conference, Proceedings, 2004*, vol. 17, pp. 39–44, doi: 10.1109/csee.2004.1276508.
- [70] T. C. Lethbridge, "What Knowledge Is Important to a Software Professional?," *Computer (Long. Beach. Calif.)*, vol. 33, no. 5, pp. 44–50, 2000, doi: 10.1007/BF03250761.
- [71] T. C. Lethbridge, "A survey of the relevance of computer science and software engineering education," in *Proceedings - 11th Conference on Software Engineering Education, CSEE and T 1998, 1998*, pp. 56–66, doi: 10.1109/CSEE.1998.658300.
- [72] B. Kitchenham, "Procedures for performing systematic reviews," *Tech. Rep. TR/SE-0401*, Keele Univ. Tech. Rep. 0400011T.1, NICTA, p. 28, 2004.
- [73] C. Spradling, L. K. Soh, and C. J. Ansoorge, "A comprehensive survey on the status of social and professional issues in United States undergraduate computer science programs and recommendations," *Comput. Sci. Educ.*, vol. 19, no. 3, pp. 137–153, 2009, doi: 10.1080/08993400903255184.

- [74] M. Goldweber et al., "Enhancing the social issues components in our computing curriculum: computing for the social good," *ACM Inroads*, vol. 2, no. 1, pp. 64–82, 2011.
- [75] S. Surakka and L. Malmi, "Need assessment of computer science and engineering graduates," *Comput. Sci. Educ.*, vol. 15, no. 2, pp. 103–121, 2005, doi: 10.1080/08993400500150762.
- [76] P. Bourque and R. E. Fairley, *SWEBOK Guide V3.0*. 2014.
- [77] H. F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," *Qual. Health Res.*, vol. 15, no. 9, pp. 1277–1288, 2005, doi: 10.1177/1049732305276687.
- [78] S. Elo and H. Kyngäs, "The qualitative content analysis process," *J. Adv. Nurs.*, vol. 62, no. 1, pp. 107–115, 2008, doi: 10.1111/j.1365-2648.2007.04569.x.
- [79] D. Gotterbarn, "Software engineering ethics," *Encycl. Softw. Eng.*, 2002.
- [80] D. L. Morgan, "Qualitative Content Analysis: A Guide to Paths not Taken," *Qual. Health Res.*, vol. 3, no. 1, pp. 112–121, 1993, doi: 10.1177/104973239300300107.
- [81] D. Bagert and N. R. Mead, "Software engineering as a professional discipline," *Comput. Sci. Educ.*, vol. 21, no. 1, pp. 73–87, 2001, doi: 10.1076/csed.11.1.73.3841.

Healthcare Logistics Optimization Framework for Efficient Supply Chain Management in Niger Delta Region of Nigeria

Imeh J. Umoren^{1*}, Ubong E. Etuk², Anietie P. Ekong³, Kingsley C. Udonyah⁴
Department of Computer Science, Akwa Ibom State University
Mkpat Enin, Akwa Ibom State, Nigeria

Abstract—Optimizing logistics allocation and utilization is essential for effective healthcare management. Apparently, less consideration is given to it in most hospitals in Nigeria where less resources are allocated to health sector in yearly budgetary. Hospital consists of several patient classes, each of which follows different treatment process flow paths over a multiphase and multidimensional requirement with scarce resources and inadequate space. Despite the small budget provision made for healthcare resources, patient's demands for better service is rapidly experiencing upsurge. Hence, efficient and optimal solutions are required to lessen costs of healthcare service towards enhancing Quality of Care (QoC) and Quality of Experience (QoE) in most public healthcare sector. However, certain control coefficients like the absence of a Dedicated Logistics Department (DLD) in the medical facilities actually limit the efforts of stakeholders. This paper proposed a Computational framework to assess various strategic and operational decisions for optimizing the multiple objectives using Type-1 Fuzzy Logic Model. In phase I, we explore healthcare resource allocation plan. In phase II, we determine a resource utilization schedule by patient class for daily operational level. While in Phase III, we develop a framework capable of evaluating and optimizing healthcare logistics using control coefficients of Logistics Optimization (LO), Integration of Information/Cognitive Technologies (ETA), and Collaboration of all Logistics Stakeholders (COL). We assigned weights between 1 and 10 to the coefficients and modeled the effects on efficient supply chain. Finally, we further explore the effects of separate strategies and their combination to identify the best possible resource supply chain. The computational experiment was considered on the basis of data obtained from a study conducted on a typical public healthcare department. Results show that our approach significantly evaluate and optimized healthcare logistics.

Keyword—Dedicated logistics department (DLD); Quality of Care (QoC); Quality of Experience (QoE); information/cognitive technologies (ETA) and type-1 fuzzy logic model

I. INTRODUCTION

The current state of Healthcare Service Delivery (HSD) in the South-South region of Nigeria—and in many other regions of the country is very precarious.

There are numerous cases where both healthcare providers or managers and healthcare users going through critical deprivation - on one hand, healthcare managers are unable to do their jobs while on the other hand, the poor masses

requiring healthcare services and other healthcare users suffer health challenges and even death. This critical challenge that have been recorded in this region can be traceable to unavailability of very required medical consumables, including medical professionals - like nurses and doctors, often resulting from many factors, one of which is lack of logistic framework for Supply Chain in the south-south, Nigeria, resulting in many deaths recorded in the emergency units of the hospitals. In one instance, University of Uyo Teaching Hospital (UUTH), many Psychiatric patients that required hospitalization were not hospitalized due to inadequate bed spaces as most hospital in Nigeria only has 0.4 psychiatric beds per 1000 population [25] Despite the fact that the South-South region of the country accounts for a great majority of the primary natural resources (crude oil) as the country exports, the region is highly neglected by the government in the provision of healthcare services. The government owned healthcare centers in the region are in many cases ill-equipped or none existent or is neglected altogether by the relevant stakeholders while most of the functioning healthcare centres are very short of personnel. In many cases, only one qualified nurse or doctor may be assigned to the centre and thus cannot handle the demands of providing quality healthcare services to the ever-increasing populace. Furthermore, due to the deplorable state of the healthcare centres,

Similarly, private healthcare centers owned by individuals are very expensive. Hence, cannot be afforded by the masses leaving in these regions who are mostly petty traders, farmers, fishermen, civil servants and the likes. The effect arising from the fact that there are no logistics evaluation framework put in place by healthcare stakeholders to address this situation. The private owned healthcare centers though functioning and equipped are still unable to meet the demands placed on them. There are still cases of running out of medical consumables because the required framework is not put in place to manage the logistics that will ensure an efficient supply chain. [7], portrays a very damning reality: Indeed, healthcare cost are increasing and healthcare managers are under increasing pressure to reduce the cost of healthcare provision - with patients expecting high quality care at affordable cost at reduced cost: there is need for better logistics practices.

A key characteristic of this region in terms of its geographical nature, is that it is swampy and thus provides enough breeding ponds for mosquitoes to thrive - as a result

the outbreak of malaria is high and the deaths recorded from malaria in this region is higher than those obtained elsewhere within the country. of the six states that constitutes South-South: Akwa Ibom, Bayelsa, Cross River, Delta, Edo, and Rivers, only Akwa Ibom State have improved impressively in sponsoring and ensuring healthcare service delivery due to the investment of the Akwa Ibom state government within the past eight years thus providing free medical care to pregnant women, children and nursing mothers and this was made possible because of proper logistics implementation. Evidently, logistics activities in healthcare centres, hospitals or clinics have been indicated to provide a significant avenue for cost containment in healthcare if best practices are implemented [7] Since businesses including healthcare industries depend on their Supply Chains to provide them with what they need to survive and thrive [11] there is validation to develop a framework that could ensure a more efficient supply chain.

II. RELATED WORKS

Not much research works has discussed the concept of healthcare logistics evaluation for the purpose of creating a more efficient supply chain for the healthcare organization. Nevertheless, many scholars have handled to a very great extent very important subjects that constitute the core of our discuss in this work. In the following paragraphs we present a summary of related works that lends credence to the subject of healthcare logistics evaluation framework for efficient supply chain. [13] established that Hospital logistics, viewed as a vital part of a hospital that is in charge of purchasing, receiving, stock management etc., accounts for up to 46% of

hospital budget—and since this amount is considered a very substantial proportion, especially in the context of budgetary restrictions applied to all organizations including hospitals and healthcare centers, they identified exhaustively logistics activities based on determining logistics manifestations within hospitals—healthcare centers. They also performed the organization and management of these activities in order to point out the departments or services, in the healthcare institutions, that handles them—this they achieved through a comparison between the various countries. They adopted a literature-review based methodology where they reviewed over 60 papers published between 2000 and 2017. From the various hospitals studied in selected countries including—France, Quebec, United States of America and Morocco, they presented the following departments and the activities they controlled or discovered as shown in Table I.

From the study summarized in Table I, we see that a singular logistics activity is handled by more than one department—this is known to create delays in the logistics process and increase the bottleneck existing in hospital logistics. Additionally, from this we discover, the diversity of logistics activities; where different departments are involved in the management of logistics activities and where some of the logistics activities are outsourced. The Administrative Department is obviously involved in too many of the activities resulting serious overhead and increased workload—delaying the time constraint of the logistics process. Noticeably, the table above showed that no Dedicated Department has been created solely for the purpose of managing healthcare logistics—a move that would have ensured efficient supply chain management!

TABLE I. LOGISTICS ACTIVITIES AND DEPARTMENTS IN CHARGE (SOURCE: [13])

Departments Logistics Activities	Medical Affairs Division	Nursing Care Division	Administrative Affairs Division	Medical Departments	Pharmacy Service	Reception and Admission Service
Scheduling	✓	✓		✓		
Procurement			✓			
Distribution			✓			
Pharmacy	✓		✓		✓	
Catering/ Food			✓			
Laundry			✓			
Hygiene			✓			
Waste Management			✓			
Maintenance			✓			
Reception Service						✓
Patient Flow						✓
Telecommunication			✓			
Information System Management						✓
Stock Management			✓			
Mail Service/ Files Archiving	✓		✓			✓
Safety and Security			✓			

In [19], in his study on the assessment of logistics management—a study carried out in Ghana Health services, ran a 3-tier system of managing medical consumables (health commodities), suppliers, the central medical store, the regional medical store, service delivery points and the transportation system. These they illustrate to form the supply chain. They adopted a multi-case study approach to assess the practices of logistics management and with these they discovered the causes of inadequacy of logistics in Ghana as well as the strengths and weaknesses of the Ghanaian Health Services' System. Many scholars have proposed numerous approaches for curbing the challenges associated with logistics management and supply chain integration. In one of these studies on Structured Review of Quantitative Models of the Pharmaceutical Supply [3] identified and provided a structured overview of quantitative models in the pharmaceutical supply chain as a means for optimizing the supply chain process. The author in [10] suggested door to door service as a means for ensuring effective healthcare supply chain based on their work on scheduling Optimization of Home Healthcare Services. They proposed that the reasonable arrangements of nurses and their routes will not only reduce medical expenses but can also enhance patient's satisfaction. In further studies, [9] after analyzing the literatures relating to the dimensions of healthcare logistics and supply chain management, shows that the areas of hospitals interfere with the determinants of satisfaction as well as the care quality standards. Inferring from the above, it could be concluded that improving the quality-of-care services and hence efficient supply chain is dependent on the efficiency of the logistics activities within the healthcare institution that will enhance efficient management. The author in [9] adopted a methodology based on the synthesis of articles and scientific reports dealing with key words/phrases of the subject matter including quality of care, hospital logistics etc. They gathered that to achieve high and improved quality, healthcare of institutions may adopt a variety of approaches, one of which is the optimization of their logistic—and this have been proposed by several scholars as one of the most efficient ways to improve both the quality of care and a more efficient management of the Healthcare institutions [16]; [9]., establishes a link between hospital logistics and quality of care and examines how strong the impact of an effective logistics is on the delivery of high quality. A Simulation-Based Multi Objective Optimization approach for Healthcare Services Management, a novel approach was proposed for healthcare services management [20]; [10]. Specifically, the use of a simulation optimization approach was prescribed for the optimal resources allocation to wards in very big hospitals. The proposed simulation-based optimization approach is based on a discrete event simulation model reproducing the hospital services and combined with a derivative-free multi objective optimization method. The results obtained on the obstetrics ward of an Italian hospital are reported, showing the effectiveness of the new approach proposed. From the methodological point of view, the main contribution of this work is the use of a simulation optimization framework, which integrates a DES model and an optimization algorithm,

allowing studying the problem in hand as a multi objective Optimization problem. Then, the DFMO algorithm used enables to obtain an approximate Pareto set of points.

III. LOGISTICS VARIABLES FOR EFFICIENT SUPPLY CHAIN

Several works have discussed key logistics variables [7]; [13] to consider for achieving specified improvements in the healthcare institutions including less cost of healthcare service; a more efficient service delivery and efficient supply chain. A cross-section of some key logistics key variables or factors are presented in Table II.

The author in [23], examined how modularity is used for enabling value creation in managing healthcare logistics services. They applied materials logistics of four different kinds of hospitals was examined through a qualitative case study. They built a theoretical framework on literature on healthcare logistics, service modularity and value creation. From the findings, their case hospitals were discovered to have developed their material logistics independently from others when looking at the modularity of offerings, processes and organizations. Services such as assortment management, shelving and developing an information platform, have been performed in-house partly by the care personnel—although should be managed by a Dedicated Logistics Department (DLD). However, steps toward modularized and standardized solutions are now being taken in the case hospitals, including ideas about outsourcing some services. Many scholars agree on centralization as being the best approach to effect an efficient supply chain but [23] maintains that modularity offers a tool for developing logistics services inside the hospitals and increases possibilities to consider also external logistics service providers.

The author in [26] postulated that, the diagnoses of diseases are carried out by medical experts with professional experience and adequate knowledge capable of identifying diseases on clinical data of patients, but such diagnosis is found to be approximate and time-consuming. The paper proposed an enhanced approach to be considered to mitigate the time-consuming nature of disease diagnoses.

Developing a framework for evaluating healthcare logistics is a conceptual equivalence of building an actual Healthcare Management System (HMS). Hence, the qualities considered when developing a system should also be considered when developing an evaluation framework. The author in [19] presented a list of quality attributes that should be considered in creating both a framework and system. This is presented in Table III.

Logistics Optimization has been earmarked as a key driver for achieving a sustainable supply chain [12]; [6] and having a centralized warehouse has been proposed as an ideal choice for healthcare institutions. The purpose for designing a framework for the evaluation of healthcare logistics is to achieve an efficient supply chain supply chain in the south-south of Nigeria and this is achieved through logistics evaluation and optimization.

TABLE II. KEY LOGISTICS VARIABLES

Variables	Applicability	Authors
Strategic Sourcing	Logistics and Supply Chain Strategizing (Strategy)	[27]
Supply Chain Network Design		
Product design & development		
Demand Planning	Logistics and Supply Chain Operations (Operations)	
Procurement		
Inventory		
Logistics		
Quality of Services	Logistics and Supply Chain Evaluation (Evaluation)	[7]
Process Complexity		
Staff Competence/Skills		
Inventory Management	Logistics and Supply Chain Implementation (Implementation)	
Process Efficiency		
Supply Chain (SC) Integration		
Cost of Implementing		

TABLE III. CATEGORIES/SUB-CATEGORIES OF FRAMEWORK QUALITY ATTRIBUTES

SN	Categories	Sub-Categories
1.	Functional Suitability	Functional completeness, functional correctness, functional correctness,
2.	Performance Efficiency	Time behaviour, Resource Utilization, Capacity
3.	Compatibility	Co-existence, interoperability
4.	Usability	Appropriateness, recognisability, Learnability, Operability, User error protection, Accessibility
5.	Reliability	Maturity, Availability, Fault Tolerance, Recoverability
6.	Security	Confidentiality, Integrity, Non-repudiation, Accountability, Authenticity
7.	Maintainability	Modularity, Reusability, Analyzability, Modifiability, Testability, Portability, Adaptability, Installability, Replaceability
8.	Effectiveness	Value ability, Specificity, Adequately perform intended task.
9.	Efficiency	Solve intended tasks with minimal resource usage
10.	Satisfaction	Usefulness, Trust, Pleasure, Comfort
11.	Minimal Risk	Health & Safety Risk mitigation, Economic & Environmental Risk mitigation
12.	Context Coverage	Context completeness, flexibility

IV. THE COMPUTATIONAL FRAMEWORK

We present a comprehensive and in-depth study of existing frameworks within the field of healthcare Logistics and Supply Chain Management in order to identify the weaknesses and functional limitations of the system. This paper employed this approach in a bid to garner a clear insight and understanding the dynamics, functionalities and requirements as well as existing knowledge as presented by reputable scholars whose works we have reviewed, required for developing an evaluation framework. With a properly structured analyses, the questions and issues to be addressed, the strengths and weaknesses, characteristics and requirements of the existing frameworks have been identified. The sole objective of carrying out this analysis, within the scope of the subject area, is to pinpoint the quality attributes, logistics variables/drivers, supply chain drivers as well as the mode of operation of the existing frameworks and their limitations with

the view of evaluating and optimizing, and solving the problems and bottlenecks associated with healthcare logistics and supply in the south-south of Nigeria.

This section considers the methods and methodologies applied to logistics and supply chain management proposed and used by scholars whose work we have reviewed in the previous chapter. The methodology used in this research work is adequately explained. It outlines the skeletal approach and design of the proposed framework.

A. Logistics Evaluation/Optimization Framework

The past three decades have witnessed the proposition and adaptation of several problem-specify methodologies. However, before the emergence of these specified methodologies and algorithm for solving problems—that could be applied repeated to get the same results at different places and in different times, randomized controlled trials

were the most reliable method of determining effectiveness [2]. However, these randomized controlled trials were discovered to have very high cost and time constraints. Often, it would take several days or several weeks to give a solution using this methodology as several randomized trials have to be made and then the most efficient trial approach that produced the best solution is chosen. This is not always the case with these trials. In many cases, no optimal solution is found even after several weeks and this validates the need for a better framework.

The introduction of a dedicated logistics department will immediately eliminate the need for retailers thus reducing the associated costs and speeding up the process and data flow. The dedicated logistics department could hitherto interact directly with the plants (producers/manufacturers) depending on the capacity of the healthcare institution or costs constraints involved with interacting with the plants and they can also interact with the suppliers thus cutting off costs associated with interacting with middlemen or the retailers. In Fig. 1, the Dedicated Logistics Department is saddled with making the decision of which is the optimal procurement option based both on transportation costs, product quality and the time between request and delivery.

B. Supply Chain Structure/Configuration

Supply Chain has been integrated into the healthcare organization as a new way approach to conceptualizing medical procurement, inventory and supply management [14]. Healthcare Supply Chain is seen by [14] as the information, supplies and finances involved with the acquisition and movement of goods and services from the supplier to the end user in order to enhance clinical outcomes while controlling costs. From this, we see that the key focus is to enhance clinical outcomes (better services) and controlling or reducing cost—which should be ensured with implementing an efficient supply chain through evaluation of Healthcare Logistics. The supply chain structure is a complex one to define when considered in a whole with all the components that it

incorporates however, the structure is quite simple and easy to graphically represent when conceptualized in smaller abstractions. We present a simple abstraction of the Supply Chain Structures that contribute to form the complete supply chain structure.

The simple supply chain abstraction/configuration considers three very important drivers: Suppliers, Company and Customers or end users. Our focus here is the company—the healthcare institution. Once abstraction of this driver is expanded, it embodies a whole lot of other drivers which are discussed in the subsections that follow. From our study, we have identified that the supply chain structure used in the several of the healthcare centres in the Niger Delta region of Nigeria is the very Extended Supply Chain Structure, with so many players involved between the producing firms and the final costumer or ultimate consumer. This is responsible for over 78% of the costs and time constraints involved in the supply chain structure used in the Niger Delta. Fig. 2 indicates a Simplified (Optimized) Supply Chain Structure.

C. Evaluation Framework Design Considerations

Based, majorly, on our literature review and the data gotten from the observation of the Medical facilities selected for our case study, we present a Healthcare Logistics Evaluation Framework (HLEF) for and propose the enactment of a *Dedicated Logistic Department (DLD)* in all healthcare facilities for Efficient Supply Chain Management (ESCM) in all Healthcare Institutions. The methodologies for the design and operation of the Supply Chain Network or Logistics Evaluation Framework is either steady state, dynamic (for which a computational intelligence approach is employed), deterministic or such that it could deal with uncertainties associated with demands for healthcare service or medical consumables. Considering that we are integrating Continuous Improvement (CI) as a design consideration for our proposed framework, a dynamic model capable of dealing with the uncertainties that may arise in future within the healthcare landscape has been adopted using Fuzzy Logic.

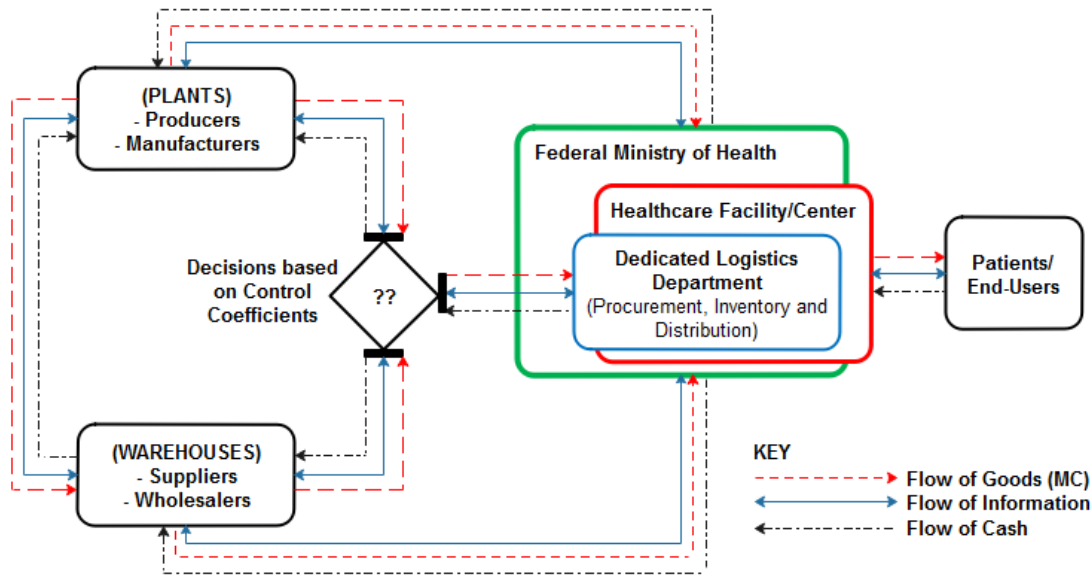


Fig. 1. Our Enhanced Healthcare Logistics Network Configuration.

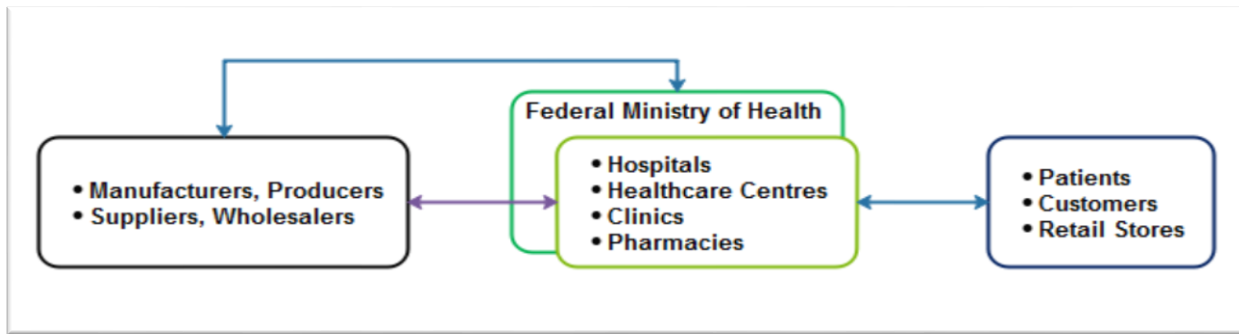


Fig. 2. A Simplified (Optimized) Supply Chain Structure (Adapted [11]).

D. Fuzzy based Framework

To development a robust Healthcare Logistics Evaluation and Optimization framework for efficient Supply Chain, we employ the combination of a multi-criteria and computational intelligence approach using Fuzzy Logic Type-1 to design a framework that implements a Dedicated Logistics Department (DLD). The Multiple criteria Technique presents modalities for implementing efficient supply chain based chosen key logistics variables (as presented in Table IV). Importantly, the absence of a Dedicated Logistics Department (DLD) and Lack of Logistics Optimization Strategy has been identified as the major control coefficients with the Highest Impact factors, using the Six Sigma framework, on the Efficiency Index of the Healthcare Supply Chain. Combining these two robust techniques, we present intelligent evaluation and optimization framework for healthcare logistics targeted at realizing an efficient supply chain. Based on this premise, we defined our fuzzy rules based on the following *if-then - else* Fuzzy rule implementation algorithm.

1) 3.41 Fuzzy rules definition: To implement the optimization module of our evaluation and optimization framework, we have defined the rules following laid down fuzzy rules. The form of a fuzzy rule is defined as a conditional statement. The fuzzy rules are defined using the standard;

$$R^1: IF x_1 \text{ is } F_1^i \text{ and } \dots x_p \text{ is } F_p^l \text{ THEN } y \text{ is } G_1^l$$

Where $l = 1, \dots, M$

From our structured analysis, we have identified (4) input variables and each of these variables have three (3) sets each. From combinational logic, we understand that a truth table of N inputs contains 2^N rows, one for each possible value of the inputs. From the 4 input variables, the maximum possible number of rules to be used in defining our rule base is given as $3^4 = 81$. As stated earlier, the rule is a collection of *IF – THEN* statements. The Linguistic Variables are shown in Table IV and the *IF – THEN* rules are shown in the Table V.

TABLE IV. LINGUISTIC VARIABLES

SN	Linguistics Variables	Key
1.	Logistics Optimization	LO
2.	Information/Cognitive Technology Adaptability	ETA
3.	Strategic collaboration of all Stakeholders and Suppliers	COL
4.	Implementation of Dedicated Logistic Department	DLD

IF – THEN RULES

If

All four parameters are at their highest *then* SCE is Efficient

Else if

DLD is adopted *and* at least any two other key variable is at their highest

Then SCE is Efficient

Else if

DLD is adopted *and* any other key variable is at its highest value

Then SCE is average

Else if

any three key variables are at its highest value

Then SCE is average

Else

SCE is Not Efficient

TABLE V. RULE BASE

Rule No.	LO	ETA	COL	DLD	SCE
	Poor	None	High	None	Inefficient
	Poor	None	Medium	None	Inefficient
	Poor	None	Low	None	Inefficient
	Poor	Minimal	High	None	Inefficient
	Poor	Minimal	Medium	None	Inefficient
	Poor	Minimal	Low	None	Inefficient
	Poor	Maximal	High	None	Inefficient
	Poor	Maximal	Medium	None	Inefficient
	Poor	Maximal	Low	None	Inefficient
	Good	None	High	None	Average
	Good	None	Medium	None	Inefficient
	Good	None	Low	None	Inefficient
	Good	Minimal	High	None	Average
	Good	Minimal	Medium	None	Average
	Good	Minimal	Low	None	Inefficient
	Good	Maximal	High	None	Average
	Good	Maximal	Medium	None	Average
	Good	Maximal	Low	None	Inefficient
	Excellent	None	High	None	Inefficient
	Excellent	None	Medium	None	Inefficient
	Excellent	None	Low	None	Inefficient
	Excellent	Minimal	High	None	Average
	Excellent	Minimal	Medium	None	Average
	Excellent	Minimal	Low	None	Inefficient
	Excellent	Maximal	High	None	Average
	Excellent	Maximal	Medium	None	Average
	Excellent	Maximal	Low	None	Inefficient

To implement the rules in Table V, we used the MATLAB Type-1 Fuzzy Logic Toolbox to realize the steps needed to actualize this framework implementation, shown in Fig. 3.

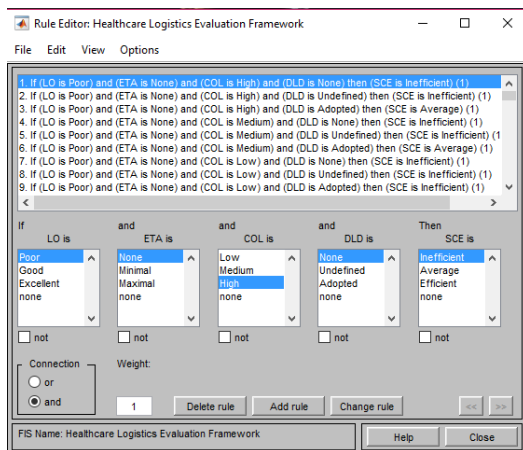


Fig. 3. Fuzzy Rule Editor.

E. Fuzzy Logic System

The conceptual system architecture used for this work is based on the Fuzzy Inference System that basically consists of the MATLAB Graphical User Interface that provides the platform for defining the Fuzzy Rules and Membership Function. The conceptual architecture illustrated in Fig. 4 consists of the Knowledge Engine, Knowledge Base which is made of the database model and the Fuzzy Logic Model and the user Interface. The knowledge engine consists of structured and unstructured data, but in this work structured data (Logistics Optimization (LO), Information/Cognitive Technology Adaptation (ETA), Strategic Collaboration of all Stakeholders and Suppliers (COL), and Implementation of Dedicated Logistics Department (DLD) are employed in the design of the fuzzy logic system. Fig. 4 illustrates the conceptual architecture of the fuzzy logic system used in the framework design.

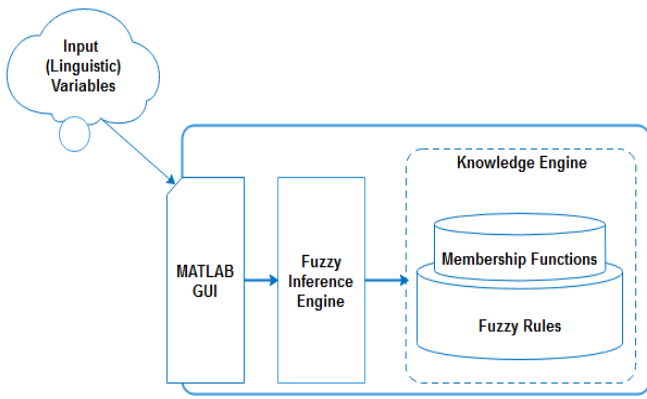


Fig. 4. Conceptual Framework for Architecture of the FIS.

F. Fuzzification

We performed a fuzzification simulation on the values of the selected input variables using the membership functions to determine their degree of membership. This converts the crisp quantities into fuzzy values. This is then used to map the output value specified in the individual rules to an intermediate output measuring fuzzy sets. The fuzzy linguistics variables and terms for each input parameter are defined as follows;

- 1) Logistics Optimization [Poor, Good, and Excellent].
- 2) Enabling Information/Cognitive Technology [None, Minimal, and Maximal].
- 3) Strategic Collaboration of all stakeholders and suppliers [Low, Medium, and High].
- 4) Implementation of Dedicated Logistic Department [None, Undefined, Adopted].

The output fuzzy linguistic variable and its terms are defined as Supply Chain Efficiency [Inefficient, Average, and Efficient]. The universe of discourse for the input and output parameters are all defined using a scale of impact significance thus;

- 1) Logistics Optimization [0, 10].
- 2) Information/Cognitive Technology Adaptability [0, 10].
- 3) Strategic Collaboration of Stakeholders and Suppliers [0, 10].
- 4) Implementation of a Dedicated Logistics Department [0, 10].
- 5) Supply Chain Efficiency [0, 10], respectively.

The crisp input and output values are converted to fuzzy values by the input and output Membership Functions (MFs) respectively. A Triangular membership functions (MFs) is used for the evaluation. A Triangular MF curve depends on three parameters b_1 ; b_2 , and b_3 , as illustrated below;

$$\mu(x) = \begin{cases} 0 & \text{if } x < b_1 \\ (x - b_1)/(b_2 - b_1) & \text{if } b_1 \leq x < b_2 \\ (b_3 - x)/(b_3 - b_2) & \text{if } b_2 \leq x < b_3 \\ 0 & \text{if } x > b_3, \end{cases} \quad (1)$$

where b_2 defines the triangular peak location, while b_1 and b_3 defines the triangular end points.

G. Membership Function Definition

In defining our membership functions, we have employed the triangular membership function in our fuzzy inference system. Individual range of inputs and output variables are outlined to relate with a fuzzy set that has the same name as the range. We have identified four Linguistic Input Variables and defined three fuzzy sets for these input variables as well as three fuzzy sets for the output variable. The Triangular Membership Functions are as defined the equations.

$$LO(x) = \begin{cases} \text{if } 0 \leq x \leq 4, | \text{"Poor"} \\ \text{if } 3 \leq x \leq 7, | \text{"Good"} \\ \text{if } 6 \leq x \leq 10, | \text{"Excellent"} \end{cases} \quad (2)$$

$$ETA(x) = \begin{cases} \text{if } 0 \leq x \leq 4, | \text{"None"} \\ \text{if } 3 \leq x \leq 7, | \text{"Minimal"} \\ \text{if } 6 \leq x \leq 10, | \text{"Maximal"} \end{cases} \quad (3)$$

$$COL(x) = \begin{cases} \text{if } 0 \leq x \leq 4, | \text{"Low"} \\ \text{if } 3 \leq x \leq 7, | \text{"Medium"} \\ \text{if } 6 \leq x \leq 10, | \text{"High"} \end{cases} \quad (4)$$

$$DLD(x) = \begin{cases} \text{if } 0 \leq x \leq 4, | \text{"None"} \\ \text{if } 3 \leq x \leq 7, | \text{"Undefined"} \\ \text{if } 6 \leq x \leq 10, | \text{"Adopted"} \end{cases} \quad (5)$$

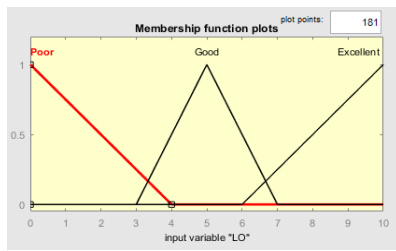
$$SCE(x) = \begin{cases} \text{if } 0 \leq x \leq 4, | \text{"Inefficient"} \\ \text{if } 3 \leq x \leq 7, | \text{"Average"} \\ \text{if } 6 \leq x \leq 10, | \text{"Efficient"} \end{cases} \quad (6)$$

The fuzzy rules are defined based on observations and analysis from our structure literature review and data collected from the case studies. Fig. 5 shows membership plots shows how the rules are applied in constructing the different membership plots.

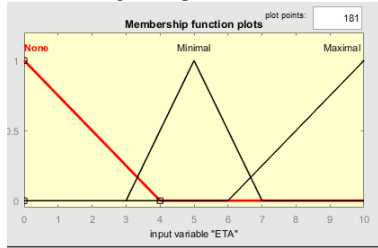
H. Fuzzy System Evaluation

The subsequent simulation results from our fuzzy evaluation were obtained as we varied the weight factors (input variables) of the control coefficients which we had initially defined on a scale where our weight factors lied between one and ten. Fig. 6 to 7 presents sections of our simulation with different input parameters.

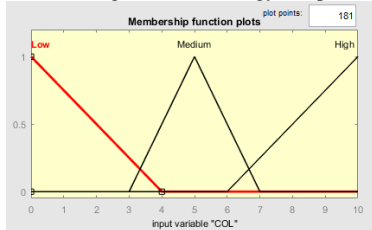
From Fig. 5(a), the input variables, LO is set at 5 which portrays Good, ETA is set at 5 which portrays Minimal, COL is set at 5 which portrays Medium and DLD is set at 5 which portrays Undefined and just as the rules define, our output: Supply Chain Efficiency is set at 5 and according to our defined rules means the Supply Chain efficiency is average. By the output being average, the system indicates that if in performing a Healthcare Logistics Evaluation, the Logistics Optimization (LO) is GOOD, the use of Information/Cognitive Technology (ETA) is MINIMAL, there is a MEDIUM Collaboration of all Stakeholders and Suppliers (COL) and the Healthcare facility's is just basic or UNDEFINED then the Supply Chain Efficiency (SCE) will be AVERAGE.



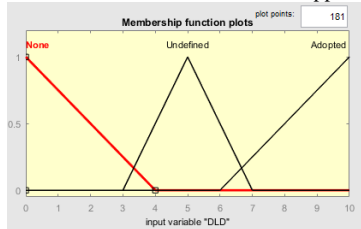
(a): Logistic Optimization (LO).



(b): Information/Cognitive Technology Adaptation (ETA).



(c): Collaboration of all Stakeholders and Suppliers (COL).



(d): Implementation of Dedicated Logistic Department (DLD).

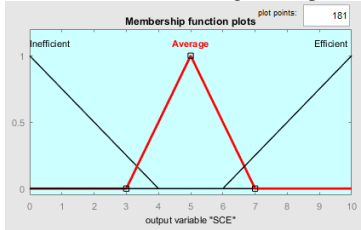


Fig. 5. (e): Supply Chain Efficiency (SCE).

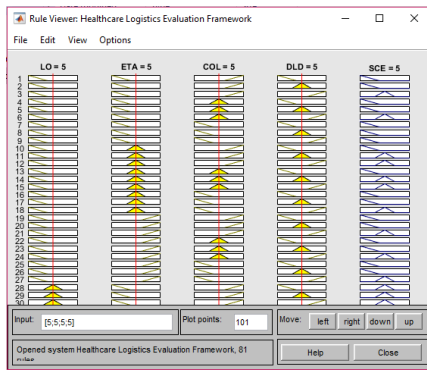


Fig. 6. First Simulation Instance.

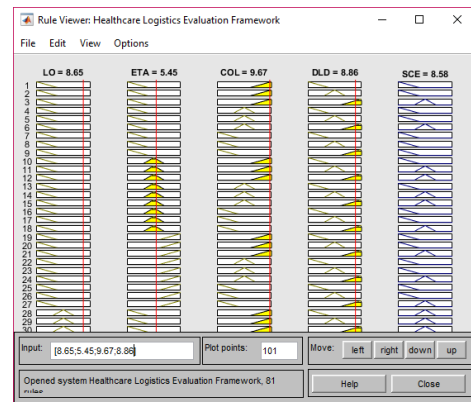


Fig. 7. Second Simulation Instance.

From our second simulation instance, the input variables, LO is set at 8.65 which portrays EXCELLENT Logistics Optimization, ETA is set at 5.45 which portrays MINIMAL usage of IT, COL is set at 9.67 which portrays a very HIGH collaboration with all stakeholders and suppliers and DLD is set at 8.86 which indicates that the hospital has ADOPTED or has a Dedicated Logistics Department and based on these inputs the system predicts that the Supply Chain Efficiency (SCE) is 8.85 which indicates that Supply Chain if implemented will be very EFFICIENT.

The surface viewer allows us to see the entire output surface of our inference system—the entire span of the output set based on the entire span of the input set. Below we present the surface plots of the input variables that impact our Logistics Evaluation Framework.

Fig. 8 presents a Surface Plot for LO (Logistic Optimization) against DLD (Implementation of a Dedicated Logistic Department). This surface plot shows that the Supply Chain is moves towards high efficiency when Logistics Optimization is EXCELLENT and the healthcare facility has a Dedicated Logistics Department to manage the Logistics affairs thus reducing the bottleneck that often arises from health workers having to combine performing logistics activities and their primary responsibilities.

Similarly in Fig. 9, COL (Collaboration of all Stakeholders and Suppliers) is plotted against ETA (Adaptation of Information/Cognitive Technologies), and from this we see that both parameters need to be at their highest (i.e. High and Maximal respectively) for the efficiency of the supply chain to move towards attaining efficiency.

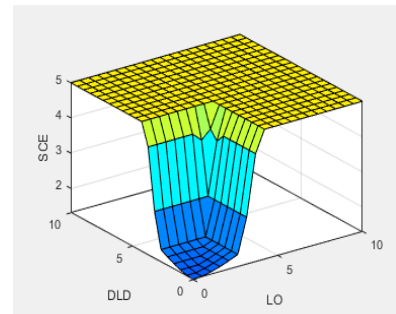


Fig. 8. Surface Plot for LO Against DLD.

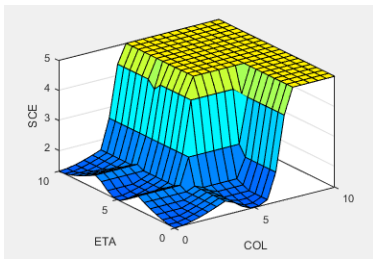


Fig. 9. Surface Plot for COL against ETA.

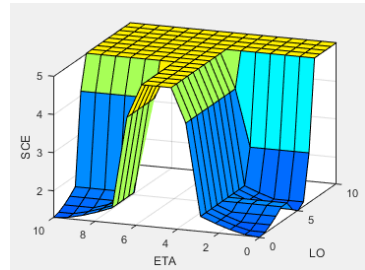


Fig. 10. Surface Plot for LO against ETA.

In Fig. 10, we consider how ETA (Adaptation of Information/Cognitive Technologies) is represented against LO (Logistics Optimization) and from the above it shows that optimizing logistics and adapting information and cognitive technologies to healthcare supply chain process will contribute to over 50% percent of the efficiency of the Supply Chain.

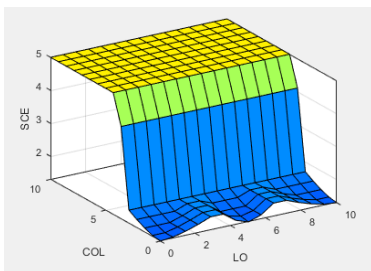


Fig. 11. Surface Plot for LO against COL.

Fig. 10 presents the impact metrics on Supply Chain Efficiency (SCE) when Logistics Processes are optimized (LO) and there is a collaboration of all Logistics stakeholders (COL). From the Fig. 11 we see that Logistics optimization has a great impact on the efficiency of the supply chain. It shows that when Hospital logistics are optimized, the more emphasis would be laid on collaborating logistics and medical stakeholders. Although this collaboration is necessary, the effect of omitting it is minimal and cannot be termed as fatal to the efficiency of our intended supply chain.

Similarly, Fig. 12 shows how the collaboration of stakeholders (COL) performs when compared against Adaptation of Information/Cognitive Technologies (ETA). Though the collaboration of Stakeholders (COL) affects the efficiency curve of the supply chain in this case, ETA affects the efficiency of the Supply Chain more than collaborating stakeholders further highlighting the importance of incorporating advancement in information technologies (IT) and cognitive science to the medical logistics and supply chain framework.

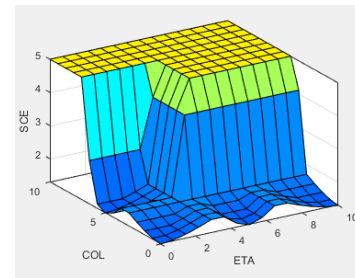


Fig. 12. Surface Plot for COL against ETA.

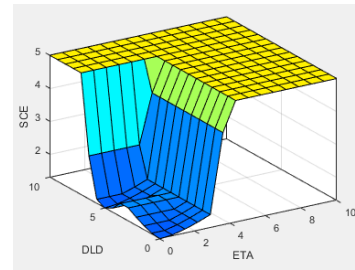


Fig. 13. Surface Plot for DLD against ETA.

Fig. 13 compares two of the variables with the very high coefficient constraints: incorporating a Dedicated Logistics Department (DLD) and Adaptation of Information and Cognitive Technologies (ETA). Both contributes similarly to the efficiency of the supply chain. These two variables are the backbone of the logistics optimization framework for efficient supply chain.

V. EVALUATION FRAMEWORK PRESENTATION

The Control Coefficients presented in Table IV presents the Decision Criteria for the design of our evaluation framework. The framework is strategically important in analysis, evaluation and optimization of Healthcare Logistics activities, processes and operations and considering that Healthcare Logistics is an integral and inseparable component of Healthcare Supply Chain therefore any improvement achieved by our Healthcare Logistics Evaluation and Optimization Framework subsequently improves the efficiency of the Healthcare Supply Chain implementation. The framework extrapolates the relationship that exists among the control coefficients as well as their bearing on the Healthcare Logistics Activities and Processes. Our framework presents a strategy for managing Healthcare Logistics Operations for the purpose of ensuring optimization and efficiency of Logistics Activities within any Healthcare Institution while effecting the realization of an Efficient Supply Chain in the Healthcare Sector.

The Healthcare Logistics Evaluation Framework presented in Fig. 14 examines and explicates the bottleneck experienced by the practice of shared management Logistics Activities by Designated Logisticians—in very few cases and Medical Personnel—Doctors and nurses, who by this shared responsibility only devote half their time and effort in performing their primary responsibilities.

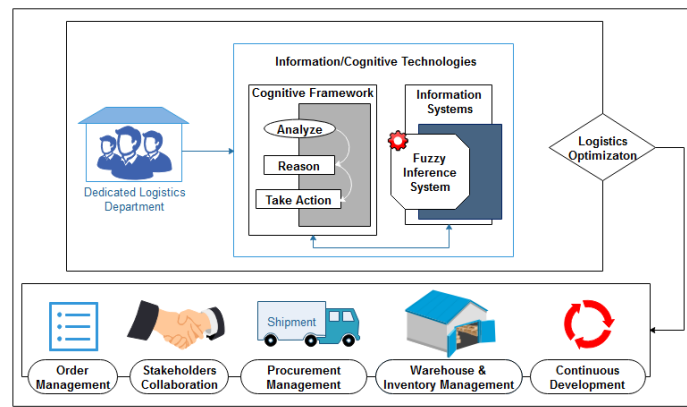


Fig. 14. Proposed Healthcare Logistics Evaluation Framework.

VI. CONCLUSION

With increasing demands on healthcare providers and hospital managers in providing Quality of Care (QoC), better Hospital resource management and allocation as well as Quality Healthcare service delivery. Indeed, the search for better approaches of managing hospital processes has increased exponentially. Given today's reality when information technologies and Intelligence Systems are the order of the day, Healthcare managers must catch up and apply cognitive approaches and top-notch techniques in information technology to tackle the problems in the healthcare industry. A very great way to begin this adoption is to better manage the thing affecting every stakeholder in the Healthcare Industry-Logistics. Considering this narrative, our framework is a timely solution to healthcare logistics evaluation and optimization for efficient supply chain.

ACKNOWLEDGMENT

Special Acknowledgement to TETFUND, Nigeria for Supporting Research and Development (R&D) in Tertiary Institutions in Nigeria.

REFERENCES

- [1] Al-Qatawneh, L., Abdallah, A. A. & Zalloum S. Z. (2019). Six sigma application in Healthcare Logistics: A framework and a case study. *Journal of Health Engineering* 2019, ID: 9691568 DOI: <https://doi.org/10.1155/2019/9691568>.
- [2] Campbell M., Fitzpatrick R., Haines A., Kinmonth A. L., Sandercock P., Spiegelhalter D., & Tyrer P. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ*, Volume 321.
- [3] Carlos F. & Alfonso-Lizarazo E. (2017). A structured review of quantitative models of the pharmaceutical supply chain; Universidad del Rosario, Escuela de Administraci ´on, Bogot ´a, Colombia.
- [4] Chen, H. K., Chen, H. Y., Wu, H. H., & Lin, W. T. (2004). TQM Implementation in a Healthcare and Pharmaceutical Logistics Organization: The Case of Zuellig Pharma in Taiwan. *Total Quality Management & Business Excellence*, 15(9-10), 1171–1178.
- [5] Denton B. T. (Ed.). (2013). *Handbook of Healthcare Operations Management: Methods and Applications*. International Series in Operations Research Management Science 184, DOI 10.1007/978-1-4614-5885-2_3. Springer Science+Business Media New York.
- [6] Dubey R., Gunasekaran A., Papadopoulos T., Childe S. J., Shihin K. T., Wamba, S. F. (2016) Sustainable supply chain management: framework and further research directions. *Journal of Cleaner Production*, 142(2): 1119-1130. ISSN 09596526.
- [7] Feibert, D. C. (2017). Improving healthcare logistics processes. DTU Management Engineering, Technical University of Denmark: PhD Thesis. Retrieved from orbit.dtu.dk on January 26, 2020.
- [8] Feibert, D. C. & Jacobsen, P. (2015). Measuring process performance within healthcare logistics—a decision tool for selecting track and trace technologies. *Academy of Strategic Management Journal*, 14. Special issue, 33–57.
- [9] Frichi Y., Jawab f., Boutahari S., Zehmed K., Moufad I., Akoudad K. & Laaraj N. (2018): Hospital Logistics; an effective tool in improving the quality of care; Laboratory of manufacturing, energy and sustainable development; Sidi Mohamed Ben Abdellah University Fez, Morocco.
- [10] Gang D., Xi Liang, & Chuanwang S. (2017). Scheduling Optimization of Home Health Care Service Considering Patients' Priorities and Time Windows. Department of Business Management, School of Business Administration, East China Normal University, Shanghai 200062, China/.
- [11] Hugos M. (2003). *Essentials of supply chain management*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- [12] Ioannis Nikolaou, Konstantinos I. Evangelinos and Stuart Allan (2013). A reverse logistics social responsibility evaluation framework based on the triple bottom line approach, *Journal of Cleaner Production* 56:173-184/.
- [13] Jawab F., Frichi Y. & Boutahari S. (2018). Hospital Logistics Activities. Proceedings of the International Conference on Industrial Engineering and Operations Management Bandung, Indonesia, 3228—3237/.
- [14] Kazemzadeh R. B., Sepehri M. M. & Jahantigh F. F. (2012). The drug logistics process: an innovative experience, *The TQM Journal* Vol. 27 No. 2, 2015pp. 214-230© Emerald Group Publishing Limited 1754-2731 DOI 10.1108/TQM-01-2015-0004/.
- [15] Kumar, A. & Rahman, S. (2014). RFID-Enabled Process Reengineering of Closed loop Supply Chains in the Healthcare Industry of Singapore. *Journal of Cleaner Production*, Elsevier Ltd, 85, 382–394.
- [16] Kriegel J., Jehle F., Dieck M. & Mallory P. (2013). Advanced services in hospital logistics in the German health service sector. *Logistics Research*, 6(2–3), 47–56.
- [17] Kriegel J., Jehle F., Dieck, M. & Tuttle-weidinger L. (2015). Optimizing patient flow in Austrian hospitals—Improvement of patient-centered care by coordinating hospital-wide patient trails. *International Journal of Healthcare Management*, 8(2), 89–99.
- [18] Manso J. F., Annan J. & Anane S. S. (2017). Assessment of Logistics Management in Ghana Health Service. *International Journal of Business and Social Research* 3(8), 75-87.
- [19] Memon, Z., Noran, O. & Bernus, P. (2019). A framework to evaluate architectural solutions for ubiquitous patient identification in health information systems. Proceedings of the 21st International Conference on Enterprise Information Systems, 580-587. ISBN: 978-989-758-372-8.
- [20] Najafi M., Eshghi K., & Dullaert, W. (2013). A multi-objective robust optimization model for logistics planning in the earthquake response phase. *Transportation Research Part E*, 49(1), 217-249.

- [21] Ohagim I. P., Nyong E. E. & Moses A. E. (2018). Methicillin-resistant staphylococcus aureus nasal carriage among surgical patients, patient relatives and healthcare workers in a teaching hospital in Uyo, Southsouth Nigeria. *Journal of Advances in Microbiology* 8(1): 1-11.
- [22] Pinna R., Carrus P. P. & Marras F. (2015). The drug logistics process: an innovative experience. *The TQM Journal*, 27(2), 214–230.
- [23] Pohjosenperä, T., Kekkonen, P., Pekkarinen, S., & Juga, J. (2019). Service modularity in managing healthcare logistics. *The International Journal of Logistics Management*, 30(1), 74-194.
- [24] Umego C. F., Mbotto C. I., Mbim E. N., Edet U. O., George U. E. & Tarh J. E. (2018). Epidemiology of hepatitis B virus infection in South-South, Nigeria: a review. *International STD Research & Reviews* 7(1): 1-17.
- [25] Umoren I., Udonyah K. & Isong E. (2019). Computational Intelligence Framework for Length of Stay Prediction in Emergency Healthcare Services Department. *IEEE Xplore Digital Library*, 2473-9464. DOI: 10.1109/ICCSE.2019.8845332.
- [26] Umoren I., Usua G. and Osang F. (2019). Analytic Medical Process for Ophthalmic Pathologies Using Fuzzy C-Mean Algorithm, *Article Innovations in Systems and Software Engineering*.
- [27] Wang G., Gunasekaran A., Ngai E. W. T. & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, 176: 98-110. ISSN 09255273.

A Data Science Framework for Data Quality Assessment and Inconsistency Detection

Anusuya Ramasamy¹, Berhanu Sisay², Amanuel Bahiru³
Faculty of Computing and Software Engineering
Arbaminch University
Ethiopia

Abstract—The accurate analysis of data requires high-quality data. However, inconsistencies occur frequently in the actual data and lead to untrustworthy decisions in the downstream data analysis pipeline. In this research, we examine the problem of the detection of incoherence and the repair of the OMD data model (OMD). We propose a framework for data quality evaluation and an OMD repair framework. We formally define a weight-based semantile repair by deletion and have an automated weight generation system that takes into account multiple input criteria. We use multi-criteria decisions based on the correlation, contrast and conflict between multiple criteria that are often necessary in the field of data cleaning. After weight generation, we present a Min-Sum dynamic programming algorithm to find the minimum weight solution. Then we apply evolutionary optimisation techniques and use medical datasets to show improved performance that is practically feasible.

Keywords—Data Science; OMD data model; weight generation; min-sum; dynamic programming algorithm

I. INTRODUCTION

Data is changing the face of the world by vitalizing creation of new drugs to fight diseases, increasing company revenues, optimization of costs, targeted advertisements or precise prediction of weather. With computers becoming increasingly powerful, high speed networks and algorithms working on vast amount of data providing competitive advantage and plethora of benefits to industry and academia. This can only be useful if the data is of desired quality; otherwise, they can be misleading or even dangerous. “Garbage in, garbage out” applies here. The quality of the input data strongly influences the quality of the results produced. In the field of data management and knowledge representation, data quality, data cleaning and consistent query answering are critical tasks but quite challenging, resulting in costly problems if not handled properly.

The concept of data quality comprises different definitions and interpretations in two main research communities: databases and management. While both communities are interested in data cleaning, the database community mostly focuses on it from a purely technical perspective whereas the management community faces the additional challenge of assessing data quality in relation to end users’ needs. In short, data quality refers to the degree to which the data adheres to a form of usage [1]. A survey listing data quality attributes that capture consumers’ perspectives on data quality showed 179 data quality attributes, which were subsequently summarized

into 20 dimensions of 4 categories: (1) accuracy, (2) relevancy, (3) representation and (4) accessibility of data [4]. In this research we focused on technical aspects of data quality of a particular format of data (known as the Ontological Multidimensional Data Models), keeping the end user in mind [5]. To ensure the quality of data, first we detect if there is any error or nonconformity and if found, we then remove the anomaly by repairing in the best possible way [14]. Normally data quality rules, such as integrity constraints, are used as a declarative way to detect errors and describe correct or legal data instances. Any subset of data which does not conform to the defined rules or constraints is considered erroneous, hence subject to repair.

The mechanism of data quality assessment and cleaning is often considered as a context-dependent activity [6]. Context can be external knowledge and/or connection to the external knowledge that confirm the validity of the given data items. Generally, context has been modelled as logic-based ontologies because of their semantic expressiveness [7]. These usually have to be expressive enough while keeping the computation complexity low, so that data extraction via query answering does not become intractable [5,13]. A database can be expressed as a logical theory, a context for it can be another logical theory and there can be a logical mapping between them to embed the database into the contextual theory or ontology. Contextual ontologies can be realized as multidimensional (MD) ontologies, due to the multidimensional nature of contexts [1,2]. These MD ontologies allow representation of dimensions as shown in the Fig. 1 the “Person” dimensional schema, which is similar to the multidimensional databases along with data tables under quality assessment. Dimensions of data are conceptual axes along which data are represented and analysed. For example, any person can have attributes which can be considered as contexts to extend knowledge about the person or verify any data involving that person. Hence, adding constraints into this system eventually supports multidimensional data quality assessment [5]. Datalog± a declarative query language (extension from plain Data log with syntactic restrictions and addition of features on the program) [8-11], has been widely used to define and extend dimension hierarchies with dimensional constraints, dimensional rules and to state formula for the quality data specifications. Dimensional rules and constraints are expressed in general syntactic forms of tuple generating dependencies (TGDs), equality generating dependencies (EGDs) or negative constraints (NCs) that extend classical integrity constraints [12].

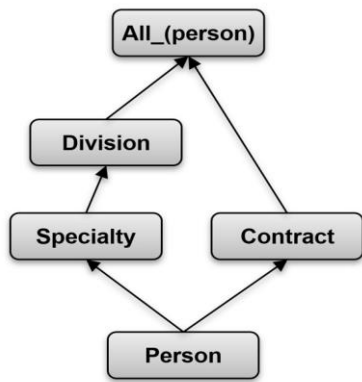


Fig. 1. "Person" Dimensional Scheme.

II. GENETIC ALGORITHMS STEPS

Genetic Algorithms (GAs) are adaptive (changes behavior at run-time) methods which are used to find the maximum or minimum of a particular function i.e. solution to optimization problems. In optimization the usual goal is to find the global optimal solution which is considered as the best solution in the whole solution space. But solution space can have obstructions associated with constraints, noise, unsteadiness, and a large number of local optima. In such situations, well designed GAs can find practically viable optimal solutions. The concept was first introduced by Holland and later it was discussed under the field of study called Evolutionary Computation where these algorithms imitate the biological process of reproduction and natural selection to solve for the fittest solutions. Just like nature, most of the genetic algorithms processes are stochastic type but efficient than random or exhaustive search algorithms.

GA begins with the population which is a set of solutions to a particular problem or objective function (Fig. 2). Each solution is usually encoded as a genotype or chromosome. If the values represented as the chromosome are continuous, those are called vectors, but if the values are just bits, those are called bit string. Ours is a discrete combinatorial problem, so we use bit string representation for the chromosomes. Each of the solutions or chromosomes is assigned a quality parameter or fitness score which measures how good the solution is to the problem. Fitness functions can also be used to differentiate

infeasible solutions from the solution space, which we also did in designing our function. The highly fit chromosomes are randomly selected for reproduction or cross-breeding which produces a child chromosome that share some features taken from each parent. In GA more than two parents are allowed but we used very basic two parent model for the crossover operation. In general, to introduce new variation in the features slight disturbance or mutation is added to the child chromosomes. This mutation basically helps against local optima and crossover explores the more promising areas of the search space. Flexible termination criteria is another benefit of using GAs. GAs also allows multiple sub-optimal solutions to be provided upon termination. The termination criteria is normally set by the user, which can be defined as number of iterations achieved, or results satisfying a given threshold. The crux here is the design of these functions. If done well the population will converge to an optimal solution to the problem.

Genetic algorithms randomly explore the whole search space and evaluate samples in many regions simultaneously, which can even be amplified by parallel computation. This strength of genetic algorithms to focus their attention on the most promising parts of a solution space is a direct outcome of their ability to combine strings containing partial solutions. In those cases, where traditional algorithms do not perform well with respect to time and space, GAs provide near-optimal practical solutions. Compared to other similar techniques like Gradient Methods, Iterated Search and Simulated Annealing, GAs offer robust and better solutions. It does not require any derivative information and performs faster and less space hungry than traditional optimization algorithms like Greedy or Dynamic Programming. It also has very good parallel capabilities. GAs work with both the continuous and discrete optimization problems, even with multi-objective functions. It provides not just single solution but set of good solutions. At any point during iteration, it has at least a solution, which is improved over time. One weakness is calculating fitness function repeatedly might be computationally expensive for some problems. In our fitness function, we incorporate satisfaction checking with Datalog queries which is also very costly operation [8-11]. Proving convergence with iterations is often not obvious and the speed at which convergence takes place is also very difficult to tackle.

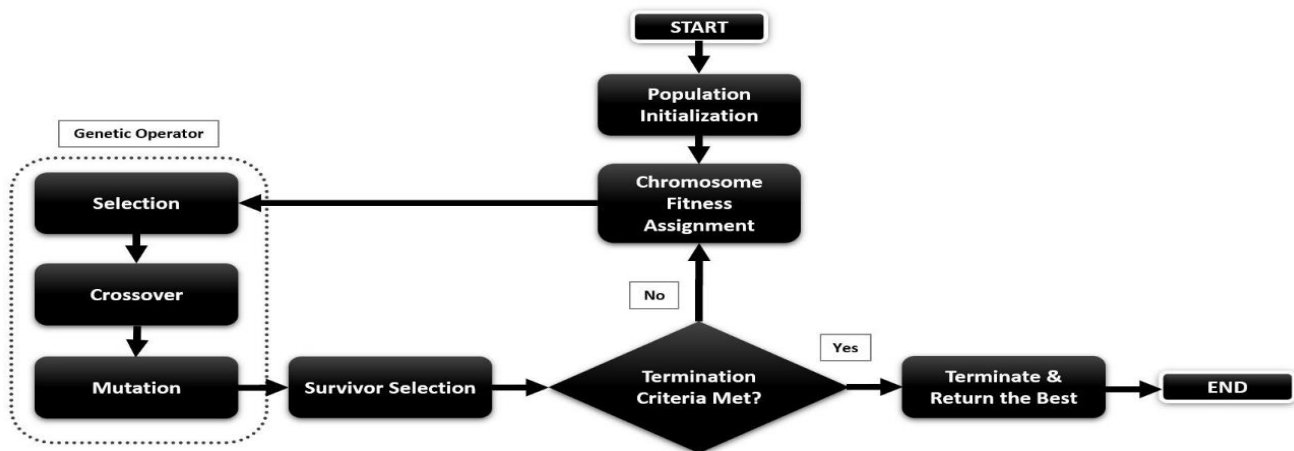


Fig. 2. Genetic Algorithm Steps.

III. A GENETIC ALGORITHM BASED APPROACH

We have a discrete optimization problem that involves selecting a subset from a set of weights that satisfies certain criteria i.e. deleting the predicates associated with those weights from the subset, will restore consistency in the OMD model. This is a bag of weights with duplicates. Not all the subsets which are subject to deletion, if deleted, will restore consistency [3]. Our total solution space has feasible and infeasible solution regions. As the first step called the initial population generation, we utilize the fitness function such that its value not only ranks each solution but also sets an outlier mark for those which are not satisfying the given constraint. As mentioned earlier about the limitations of GAs, we also cannot guarantee the convergence with iterations in our version. But we start with the chromosome containing the superset of all the candidate inconsistent predicates, which is obviously a solution in the solution space. Then, we keep on selecting chromosomes of proper subsets randomly and calculate their fitness to evaluate as a possible solution with minimal weights.

In the discrete optimization problem, Genetic Algorithms (Fig. 2) usually consider chromosomes as binary strings which consist of 1s and 0s indicating whether the indexed item is selected. For example, if the set of predicates is $\{\mu_1, \mu_2, \mu_3, \mu_4\}$ which are the sources of inconsistency in an OMD model, and the respective weights for these predicates are $\{2, 3, 4, 5\}$, one chromosome could be $[1, 0, 0, 1]$, this represents selecting predicates $\{\mu_1, \mu_4\}$ with weights $\{2, 5\}$ for deletion. In our examples below, for the sake of simplicity, we show weight set $\{2, 5\}$ format instead of the binary set $[1, 0, 0, 1]$ format.

Fitness: The fitness function takes as input the chromosome consisting of the predicates which are selected for calculating fitness, then source of inconsistency, and the set of constraints. After excluding the predicates (κ) from the set of inconsistent predicates (μ), we check whether the constraints are satisfied. If the remaining predicates ($\mu \setminus \kappa$) satisfies the constraint (η), we return the sum of weights for all predicates in κ with the weights in W . If the predicates do not satisfy the constraint, we return the infinite number to indicate this is an unfit chromosome, hence, ignore this candidate. Formally, we define our fitness function as:

$$\text{Fitness}(W) = \begin{cases} \sum_{i=0}^{|W|} w_i & (\forall w_i \in W) \\ \infty, & \text{Otherwise} \end{cases}$$

This defined fitness function terminates as we always return a value. In the case when the constraint is satisfied, we iterate through a finite set of elements bounded by the size of W , and return the sum of the predicate weights. In the case the constraint is not satisfied, we simply return a large (infinite) value.

In this algorithm, initial weight lookup and summation of weights take linear time to compute. But, the core part of the Algorithm 1 is a Datalog \pm query to check satisfiability ($\delta \eta$) which is a EXPTIME-complete problem itself. If we consider this particular query is taking τ time in the worst case, then the complexity of this algorithm can be shown as: $O(|\kappa| + |W| + \tau)$.

Algorithm 1: Fitness

Input: List of predicates under consideration, all inconsistent predicates, constraints

Output: Fitness score of the chromosome

```
1: procedure Fitness( $\kappa, \mu, \eta$ )
2:    $W \leftarrow \emptyset$  . Set of Weights
3:   for each  $m \in \kappa$  do
4:      $W \leftarrow W \cup \text{WeightAtom}(m)$ 
5:   end for
6:    $S = 0$  . Sum of all weights of the sub multiset
7:   for each  $w_i \in W$  do
8:      $S = S + w_i$ 
9:   end for
10:   $\delta \leftarrow \mu \setminus \kappa$  . Subset removed from superset
11:  if  $\delta \eta$  then . Checking consistency after deletion
12:    return  $S$ 
13:  end if
14:  return  $\infty$ 
15: end procedure
```

Algorithm 1 shows the fitness calculation. We take the predicates from the chromosome κ (i.e. potential deletion candidates) and drop those predicates from μ and check the consistency against the constraint without the remaining atoms in μ . If satisfied, we return the total weights S of all the predicates in κ . If not satisfied, we return a large integer, at least larger than sum of all the weights, to designate this chromosome into the infeasible solutions space of the population. We consider only those chromosomes, where deleting the items indexed there, will satisfy the constraint and their total weight is less than sum of all the weights of predicates in the source of inconsistency list. For example, if the weight list is like: $[1, 2, 3, 5, 10]$ and sum of these are $(1 + 2 + 3 + 5 + 10) = 21$. Given two chromosomes, say $[2, 3, 5]$ and $[1, 2, 3]$, deleting them both satisfies the constraint, then their respective fitness score is: $(2 + 3 + 5) = 10$ and $(1 + 2 + 3) = 6$. Any candidates which do not fall into this range, are assigned a score larger than 21 so that it is discarded. We consider low fitness scores as better chromosomes, as our objective is to find minimal weight.

Population: As we know that, not all the regions in the total solution space are feasible, we have to design this population initialization Algorithm 2 in a way so that it can only produce those chromosomes which are feasible. We utilize the fitness function to determine feasibility. It also takes into account the size of the population.

Algorithm 2: Population

Input: Size, total list of inconsistent predicates, constraint

Output: Priority Queue of Chromosomes

```
1: procedure Population (N,μ,η)
2:   Θ ← μ
3:   for all i ← 1...N do
4:     Max ← Top(Θ)
5:     MaxWeight ← Fitness(Max)
6:     κ ← RandomChromosome(μ)
7:     if Fitness(κ) < MaxWeight then 8: Insert(Θ,κ)
9:   end if
10: end for
11: return Θ
12: end procedure . Priority Queue
```

Algorithm 2, already contains the fitness function (Algorithm 1 with the worst-case complexity $O(|\kappa| + |W| + \tau)$). There is also a priority queue “Insert” function with the worst-case complexity of $O(\log N)$ where N is the size of the queue. As these two functions run N times to generate the population, the overall complexity of this algorithm becomes $O(N \times (|\kappa| + |W| + \tau + \log N))$.

The population function returns a max priority queue of a user-defined size. The idea behind using the priority queue, is to narrow down solution space and cross-over area. Whenever any new chromosome is generated and found to be fit, then if its sum of weights are smaller than the max in the queue, it pops out the max item and inserts the new chromosome. For example, if the max priority queue consists of weights: [10,7,3] and a new chromosome comes with the weight 15, which is feasible as it's less than $(10+7+3) = 20$. But in the max priority queue, the weight 10 is the maximum, so this candidate will not be inserted. If a new chromosome with fitness weight 5, that is less than 10, to keep the size of the queue 3, it will pop out the current max 10 and insert 5. Hence, the new priority queue will be [7,5,3].

Crossover: The crossover breeds new chromosomes which are better in quality, which means they have better fitness score i.e. of lower value. Crossover takes features from both the parent chromosomes. Here it takes the max priority queue as input and produces the new chromosome as output.

Algorithm 3: Crossover

Input: Max Priority Queue of Population

Output: New Breed Chromosome 1: procedure Crossover(PQ)

```
2:   κ1 ← RandomChromosome(PQ)
3:   κ2 ← RandomChromosome(PQ)
4:   κ ← κ1 ∩ κ2
5:   return κ
6: end procedure
```

Algorithm 3, has just two constant time random chromosome selection functions and an union of two chromosome operation which has the running time of $O(|\kappa|)$ where κ is the length of the chromosome.

The idea behind our crossover Algorithm 3 is that, those chromosomes which are of minimal weighted set of atoms, they have high probability of appearing in the super multi-sets containing them, where these super multi-sets if deleted, restores the consistency. To get minimal weights we can take the common atoms among the two parent chromosomes. For example, if the two feasible parent chromosomes are: [1,2,3,4] and [2,3,5,7,8], then there is a possibility that the multi-set with the common items [2,3], is the minimal chromosome which has better fitness $(2+3) = 5$. So, we randomly choose two parents from the max-priority queue and produce a new chromosome by selecting the common predicates between them.

Mutation: Mutation is the technique to introduce new features which may or may not be present in the parents. In our context, this is just to mutate or change some bits such that it introduces new features outside of the current domain. Our mutation algorithm takes in the new breed generated from the crossover, changes a bit if applicable and produces the new child chromosome for fitness testing and adding to the priority queue.

Algorithm 4: Mutation

Input: Chromosome, Population

Output: Child Chromosome

```
1: procedure Mutate(κ,Population)
2:   κ1 ← RandomChromosome(Population)
3:   κ2 ← RandomChromosome(Population)
4:   i ← RandomInteger(0,Length(chromosome))
5:   if κ[i] == 1 then
6:     κ[i] = (κ1 ∩ κ2)[i]
7:   end if
8:   return κ
9: end procedure
```

All the operations in the mutation function are of constant time, so the complexity of this algorithm is just $O(1)$.

In Algorithm 4, mutation works in the population's feasible solution region and selects two of the fit chromosomes randomly which are not in the priority queue. Then randomly chose one position in the child (input) and matches that position with the two randomly selected chromosomes from the population. The algorithm changes the bit in the child with the one found in the randomly selected chromosomes, if they are equal. For example, if the child is [1,2,3,4,5], and both of the randomly selected chromosomes do not have 4 in the 4th position, then we mutate the child into: [1,2,3,5] by discarding 4, and if it is a good fit, we can insert this child into the queue.

Iteration of Genetic Algorithm: Finally we implement the iteration phase of the genetic algorithm, where the crossover and mutation continue running until it converges to a point where no other improvement is observed. Other termination criteria such as number of iterations or solutions or even fixed running time can also be used. The benefit of using GA is, at any iteration, there is a solution available. Although it may not be the optimal one, over subsequent iterations, the solution improves.

IV. RESULTS AND ANALYSIS

This research is the first step towards inconsistency restoration of ontology multi-dimensional data models. This is also based on Datalog±, for which there are not many matured tools or libraries readily available. So, we developed a working prototype and synthetic datasets to test our algorithms. Our objective here is to discuss about the system used for implementation.

A. System Configuration

We ran our experiments using virtualization of the Linux server (Architecture x86-64) with 32GB RAM, running Linux Mint 19.1 operating system on Intel Xeon (CPU E52687W v4 @ 3.00GHz). All of the implementations were done in the Python programming language, except weight generation algorithm which was done in R. To simulate the Datalog± behavior, we used python’s plain Datalog library known as pyDatalog.

B. Data Set

Our datasets are the based on the running example we created but much larger in size to resemble practical usage. The

information to enrich dimensions are also inspired from real world knowledge bases.

We introduced two dimensions “Person” and “Drug” and their dimensional instances as a toy example (Fig. 3). Here, we have kept the same dimensional schema but extended dimensional instance, indicating the number of instances in each category by a circled integer (Fig. 3).

“Person” has 2 Divisions, "Doctor" and "Nurse". "Doctor" has 7 specializations and "Nurse" has 3, which is in total 10 specialties, therefore circled 10 displayed beside category “Speciality” in the Fig. 3. Specialities of the Doctors are: "Cardiologist", "Pediatrician", "Medicine", "Gynecologist", "Surgeon", "Dermatologist", "and Neurologist". Specialities of the Nurses are: "Clinical", "Forensic", "Orthopedic". For the “contract”, it can be of "Full-time" or "Intern" i.e. 2 types of contracts.

At the bottom, we have “Person” category containing names of the 100 doctors or nurses. Drugs are of 2 types "Restricted" and "General Sale". There are total 16 drugs are in “Drug” category; 3 of them are listed as "Restricted" type and 13 of them are of "General Sale" type. Restricted drugs are: "Santonin", "Meclozine", "Ketamine" and general sale drugs are: "Ibuprofen", "Plasmin", "Carprofen", "Histamine", "Lipitor", "Nexium", "Plavix", "Abilify", "Seroquel", "Singulair", "Crestor", "Actos", "Epogen".

For the data tables, “Administer Drug” and “Bills” the same rule and schemas were kept, but we generated different sizes from 10 to millions of records (Fig. 4) for testing the algorithms developed for restoring consistency with the dimensions built above.

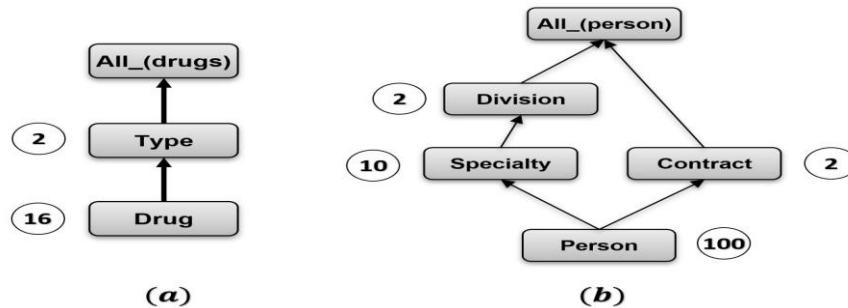


Fig. 3. Drug (a) and Person (b) Dimension (with # of Instances).

AdmDrug (t, pb, d; p, ag)

Time	Prescribed By	Drug	Patient	Age
28-Mar-18	Emdad	Ibuprofen	Rafi	80
12-Feb-18	Tom	Plasmin	Anika	2
14-Feb-18	Tom	Santonin	Ruby	1
16-Dec-17	David	Santonin	Harry	15

Bills (t, sp, dt; p, am)

Day	Specialization	Drug Type	Patient	Amount
28-Mar-18	Cardiologist	General Sale	Rafi	200
12-Feb-18	Pediatrician	General Sale	Anika	150
14-Feb-18	Pediatrician	Restricted	Ruby	60
16-Dec-17	Clinical	Restricted	Harry	?

σ

10 ... 1 Million Tuples

Fig. 4. Dataset Tables (with # of Instances).

C. Source of Inconsistency

This includes detection and searching the ground atoms which are responsible for the inconsistency. This part is developed using “pyDatalog” library and search procedure expressed in the form of query answering. As we can see in the Fig. 5, it is almost linear in nature, that means, the time (seconds) required to get all the ground predicates is proportional to the number of records in the dataset.

D. Weight Generation

We have introduced 6 criteria and after obtaining the deletion candidate predicates, we arranged them in a matrix and used the CRITIC method to generate the weights. Fig. 6 shows the runtime (in milli-seconds) as we scale the number of predicates. We used the R language for this purpose. All the steps in CRITIC method are mathematical functions operating on the single matrix, so the calculations are very fast and scalable. For 1,000,000 predicates it took only 2.5 seconds to generate the weights.

E. Deletion Candidate Search

After receiving all the predicates, identified as the sources of inconsistency, we execute Genetic Algorithm, which computes the deletion candidates with minimal weight sum to the end user. Fig. 7 displays the search space, number of tuples with subsets and performance (time), in the single graph to help us easily comprehend their interrelationship associated. It shows the running times (right Y-Axis) on varying input size and also how many of the predicates (in percentage, left Y-Axis) are being considered as deletion candidates among the total inconsistent predicates. Here, X-Axis shows different input sizes, that means number of tuples in the fact tables. Inside the parenthesis, it shows the total number of subsets required to generate in the worst case. For example, the 2nd data point, $n = 50(16)$, expresses that, there were 50 records and out of them 4 were found as sources of inconsistency i.e. number of subsets to generate at worst case was 24 or 16.

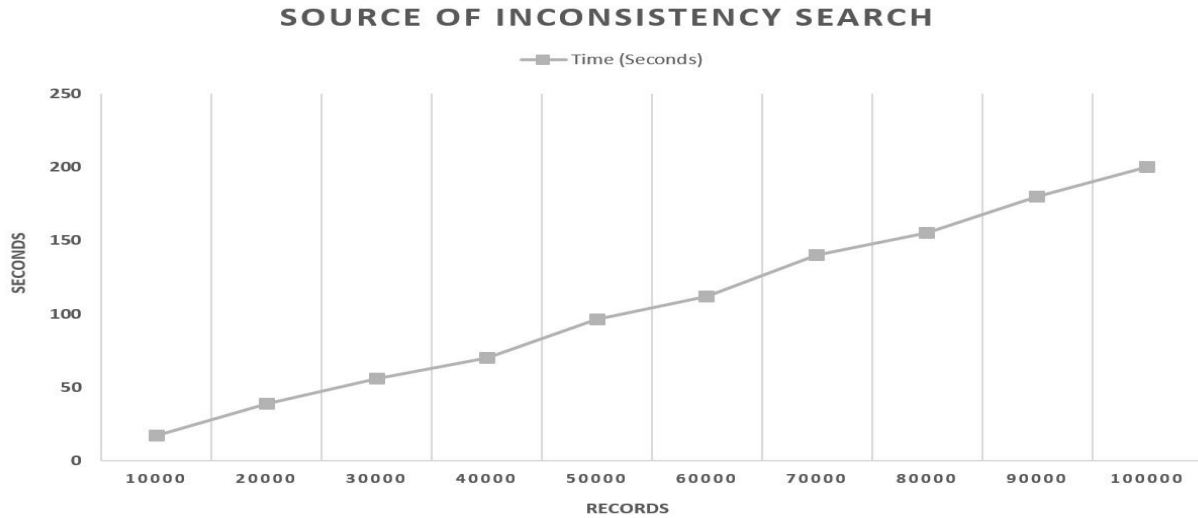


Fig. 5. Source of Inconsistency Search Performance.

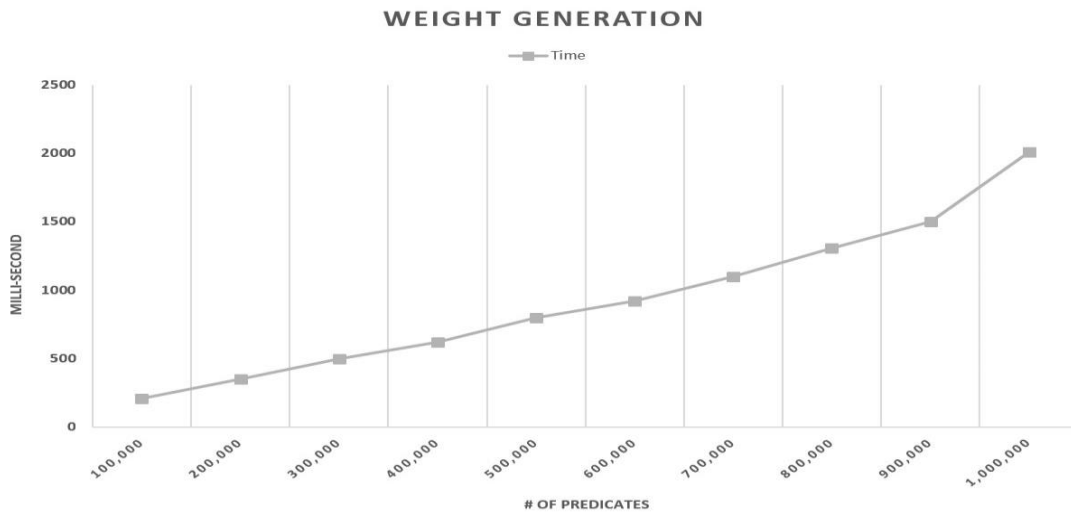


Fig. 6. Weight Generation.

COMBINATION GENERATOR FOR SUB-MULTISET WEIGHTS SUM

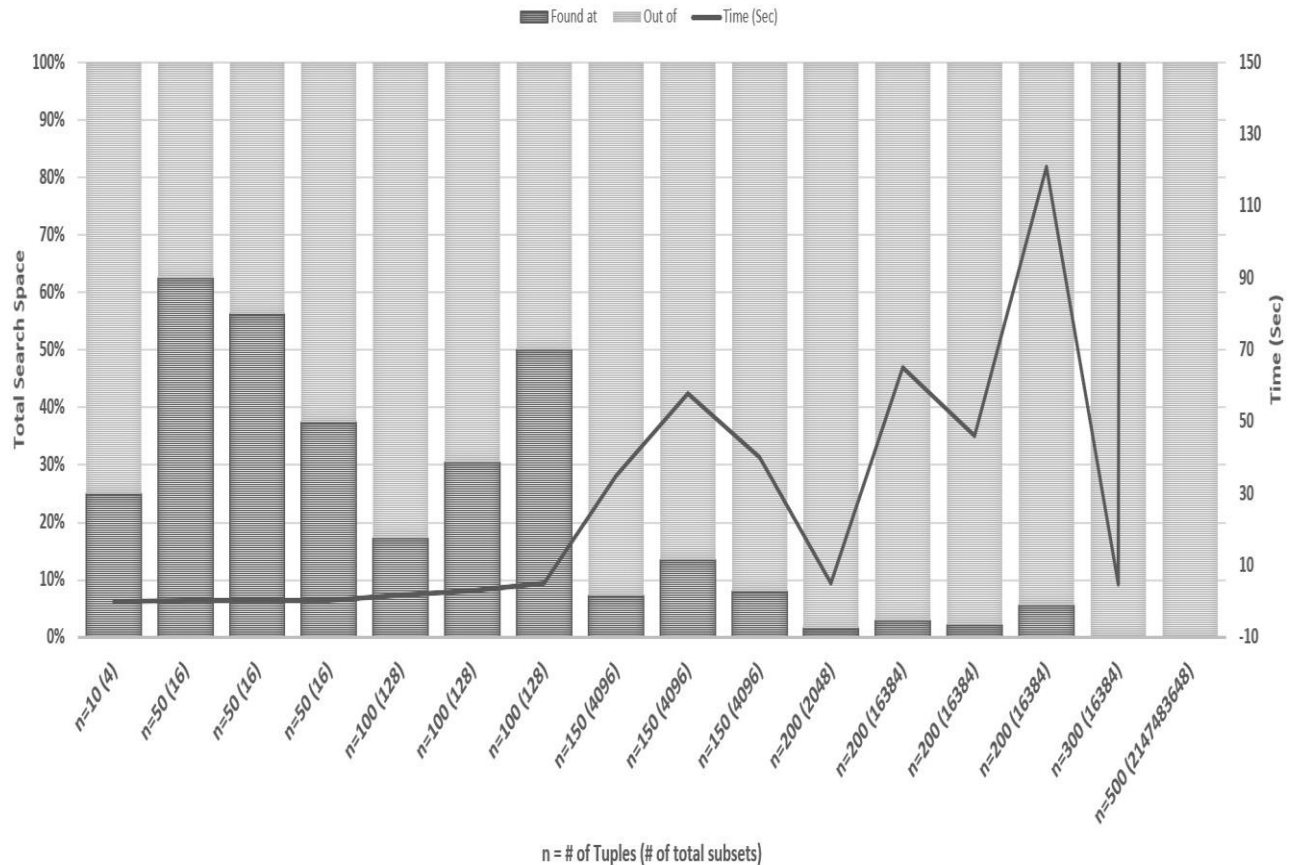


Fig. 7. Greedy and DP based Algorithm Performance.

As it was mentioned earlier that, the idea behind developing this Greedy-DP based algorithm was to generate combinations of sub-multisets in the ascending order of their sums, so that all the models or sets were not required to generate which could be exponentially growing. This graph (Fig. 8) is actually empirical evidence that not all the models need to be generated. In our experiment, in all of the cases, out of all the models i.e. out of 100% (lighter part), around 20% (darker parts) were required to generate to get the minimal weight. Fig. 9, shows that time is proportional to the number of subsets generated. Time performance measurement is actually trivial, because even though there could be 1 million records but the first smallest weight could be the only one deletion candidate and it would take less than a second to find it, whereas with 500 records only, if the deletion candidate is far away from the minimum weight, it could take longer time to find the expected subset of minimal weight.

To solve the worst case scalability problem with the Greedy-DP based algorithm, we trade off guaranteed minimum weight and utilize sub-optimal genetic algorithms for better performance with respect to time and space. After following

the steps described in the Algorithms-(1,2,3,4) we found better results. For example, with the designed dimensions and fact-tables of 1500 records (AdmDrug and Bills), which had 80 source of inconsistency ground atoms (with $280 = 1,208,925,819,614,629,174,706,176$ Models), The population ran for 1 hour 45 minutes and then crossover-mutation ran for another 15 minutes, in total 2416 iterations in 2 hours resulted in exact minimum solution whereas same problem took more than 3 days to be solved by the fastest optimum Greedy-DP based solution.

As we can see in the Fig. 10, at first the population is initialized by taking all the items i.e. sum of all the weights (12401) of the 80 inconsistent predicates, then, converged slowly towards lower minimal weights. In the iteration phase, the result kept on improving with crossover and mutation, which sharply converged towards the desired solution (sum of weights 163). The practical benefit of using this algorithm is, a user can still run the iteration phase and at any time when the algorithm stops, a minimum weight subset solution is generated up to that time. Deleting such candidates would enable consistency in the OMD model.

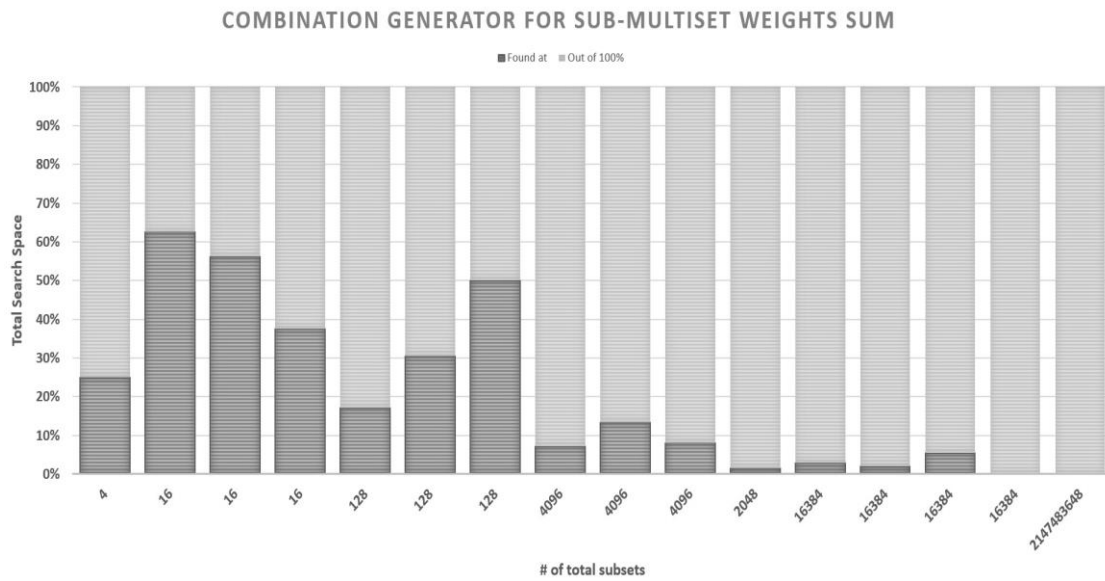


Fig. 8. Total Number of Set Generated to Find Solution in Search Space.

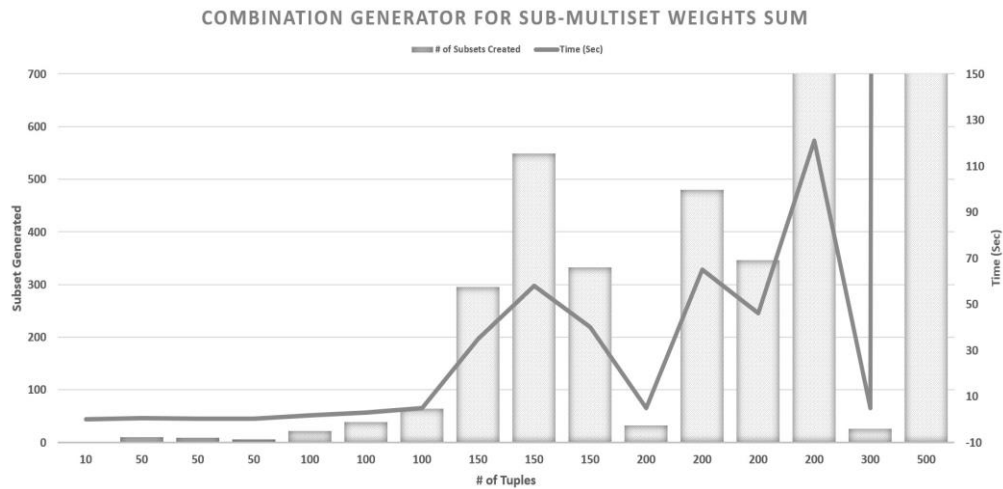


Fig. 9. Subsets, Tuples and Time Relation.

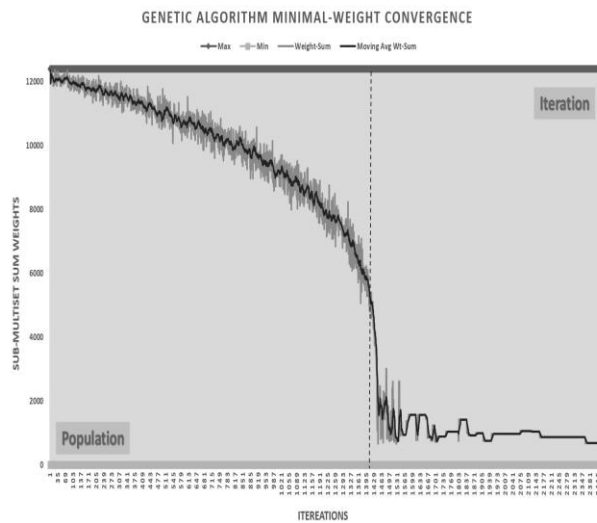


Fig. 10. Genetic Algorithm Iterations for Minimal Weight Search.

V. CONCLUSION

We studied the inconsistency detection and repair problem for the OMD model with respect to a set of dimensional constraints and rules. We showed how rules and constraints complicate the repair generation process. We presented our technique to detect inconsistencies in the tuples or predicates of the dimensions, and proposed a weight based repairing algorithm to restore consistency in OMD models. Given the multiple criteria that may be needed to generate an objective set of weights, we used the CRITIC method to compute a set of weights based on multiple criteria decision making, without user intervention. We also formally defined the minimal-weight repair semantics for the OMD model, and presented algorithms to identify the source of inconsistencies and to ground the generated predicates. We then developed a greedy and dynamic programming based minimal weight searching algorithm which outputs the predicates of minimal weight as final deletion candidates. The idea behind this algorithm was that, we did not need to generate all the models of a given theory to get minimal weights, if we could generate models in the ascending order of their sum of weights, that would be sufficient assuming that on an average, the expected set of predicates would be found at the midpoint of the search procedure. Our evaluation showed that, our assumption was correct about Min-Sum algorithm. This approach is faster than using brute force technique. We also implemented Genetic Algorithms for practical usage, which could be sub-optimal, however, our experiments demonstrated superior performance in terms time and space.

This research encompasses consistent query answering, ontologies, data cleaning, number theory and evolutionary algorithms. However, we see three avenues for future research: We have proposed deletion based weighted repair, but update-based weighted repairs would be more interesting to explore. From the perspective of mathematical logic and SAT solvers, to investigate if there is a way to generate models in ascending order of their summation of weights. This would eliminate costly satisfaction checking in the Min-Sum algorithm. There is prior research on updating dimensions. However, dimensions can be replaced with graphs or other types of ontologies in combination with tuples. OMD model has structural constraints, if those are relaxed, it can be a more expressive

ontology or more general like graphs. The techniques used here i.e. weights generation and finding their minimal subset using Min-Sum and GAs, are generic in nature. So, the application of such algorithms can be extended to diverse ontologies or graph databases.

REFERENCES

- [1] E. Haque and F. Chiang, "Restoring consistency in ontological multidimensional data models via weighted repairs", *Procedia Computer Science*, vol. 159, pp. 1085–1094, 2019.
- [2] L. Bertossi and M. Milani, "Ontological multidimensional data models and contextual data quality", *Journal of Data and Information Quality (JDIQ)*, vol. 9, no. 3, p. 14, 2018.
- [3] R. Janicki, "Finding consistent weights assignment with combined pairwise comparisons", *International Journal of Management and Decision Making*, vol. 17, no. 3, pp. 322–347, 2018.
- [4] C. Batini and M. Scannapieco, "Data and information quality: Concepts", *Methodologies and Techniques*. Switzerland: Springer International Publishing, 2016.
- [5] A. Arioua, N. Tamani, and M. Croitoru, "Query answering explanation in inconsistent Datalog+/- knowledge bases", in *International Conference on Database and Expert Systems Applications*, Springer, pp. 203–219, 2015.
- [6] I. F. Ilyas, X. Chu, et al., "Trends in cleaning relational data: Consistency and deduplication", *Foundations and Trends® in Databases*, vol. 5, no. 4, pp. 281–393, 2015.
- [7] R. Estrella, D. Cattrysse, and J. Van Orshoven, "Comparison of three ideal point based multi-criteria decision methods for afforestation planning", *Forests*, vol. 5, no. 12, pp. 3222–3240, 2014.
- [8] L. Bertossi, F. Rizzolo, and L. Jiang, "Data quality is context dependent", in *International Workshop on Business Intelligence for the Real-Time Enterprise*, Springer, pp. 52–67, 2010.
- [9] S. Ceri, G. Gottlob, and L. Tanca, *Logic programming and databases*. Springer Science & Business Media, 2012.
- [10] G. Orsi and L. Tanca, "Context modelling and context-aware querying", in *International Datalog 2.0 Workshop*, Springer, pp. 225–244, 2010.
- [11] A. Cali, G. Gottlob, T. Lukasiewicz, B. Marnette, and A. Pieris, "Datalog+/-: A family of logical knowledge representation and query languages for new applications", in *2010 25th Annual IEEE Symposium on Logic in Computer Science*, IEEE, pp. 228–242, 2010.
- [12] C. A. Hurtado, C. Gutierrez, and A. O. Mendelzon, "Capturing summarizability with integrity constraints in olap", *ACM Transactions on Database Systems (TODS)*, vol. 30, no. 3, pp. 854–886, 2005.
- [13] L. Bertossi and J. Chomicki, "Query answering in inconsistent databases", in *Logics for emerging applications of databases*, Springer, pp. 43–83, 2004.
- [14] W. W. Eckerson, "Data quality and the bottom line", *TDWI Report*, The Data Warehouse Institute, pp. 1–32, 2002.

Investigation of Smart Home Security and Privacy: Consumer Perception in Saudi Arabia

Omar Almutairi¹

Computer Science Department
Shaqra University, Shaqra
Saudi Arabia

Khalid Almarhabi²

Department of Computer Science, College of Computing in
Al-Qunfudah, Umm Al-Qura University
Makkah, Saudi Arabia

Abstract—One of the fastest and most developing technologies around the globe is the Internet of things (IoT). The research questions in this study focus on the security and privacy challenges for a smart home environment. The geographical region of Saudi Arabia is the selected boundary for the study. The study is focused on finding the problems associated with the Smart Home adaption in Saudi Arabia. However, there is a large phase shift, which is seen towards the increase of threats in smart homes. It is believed that the awareness by humans towards the use of these devices. The level of security offered by the devices, is one of the factors for these threats and privacy issues. This research targets to identify all the facts that can be discarded towards adaption of Smart Homes. It is desirable that a quantitative methodology must be implemented for identification of the population under threat due to IoT devices in smart homes. The views of the users are the major input values to trace the problems. The expected results from this research will provide all the factors which can be improved and provided with proper solution to avoid any security or privacy threats in the Saudi Arabian realm.

Keywords—Smart home; IoT; Saudi Arabia; security; privacy; issues; demographic; perception; consumer

I. INTRODUCTION

The Internet of things (IoT) refers to the connection of physical objects that are created with the help of sensors and connected to a local area network. The devices exchange data with the centralized database server systems, which in return are capable of processing the information passed on from these devices. The decision-making and the business flow take place with the help of these decisions and the analysis done by the combination of the devices and servers. The evolution of such devices started in 1982, when a Coca-Cola vending machine was able to send information about skates inventory and sales over the Internet at Carnegie Mellon University [1].

The ecosystem of the IoT-based devices is comprised of five basic elements [2]. First and foremost are the sensing and embedding components. These devices are loaded with any kind of sensor to provide specific functionality. Second, another class of elements is the connectivity and networking components that empower the devices to communicate with a centralized communication unit, or with a server, more specifically. Improvement in the use of IOT leads to adaption of cloud technologies. The IOT Cloud becomes the third important element in this context. The usage and the information for any device that is empowered with IOT-based

learning is recorded in the cloud. The management of the data is done in the cloud. However, the analysis is done with the help of sophisticated analytics and data management services. The endpoint of the realm is comprised of the end user devices that have an interface for communication.

There is a strong change observed towards the digital environment in today's challenging world. Practically everything is sensed and recorded to derive conclusions and define business processes [3]. The integration of all the processes is done with the help of Internet- and sensor-based devices. It is really an important improvement, which is required at this point in time. However, it also has its own drawbacks. There is a strong requirement of trusted frameworks for the smooth working of this drastic digital upgradation. According to the Vision 2030 of Saudi Arabia, the adaption of digitization is one of the major factors. The use of Internet and communications technology to improve the quality of living and to facilitate the citizens is indeed one of the most important aspects towards use of IOT [4]. Various projects going under the National Committee for Digital Transformation in Saudi Arabia since 2006 have started achieving national digitization. The most promising factor for this relates to digital health, education, e-commerce and smart cities, including smart homes. A promising study conducted by [5] reveals the fact that even the citizens as well as the residents are looking forward to improvements in the quality of the standard of living, with the use of smart homes and IOT-based devices.

II. THE SMART HOME ENVIRONMENT

The general thinking that arises in the mind about smart homes is that they are comprised of security and surveillance systems. However, these systems are not enough to make a home smart. A taxonomy of the smart home services was presented by [6]. The main categories in which a smart home can be expected to work are comprised of detection of health conditions, storing and retrieving multimedia information, security and surveillance, and, finally, device monitoring for energy conservation. The broader classification of a smart home can include safety, energy consumption management, and lifestyle support. The ultimate global extinction of nonrenewable resources of energy has led to the necessity of identifying renewable sources of energy and reducing the usage of energy consumption. Smart homes are providing a new era for the conservation of energy and supporting the cost of living of humanity.

The very first smart home device of its kind was developed in 1966–67, which was called ECHO-IV. The device was capable of managing shopping lists, controlling the temperature of the house and turning off certain appliances [7]. However, the device was not sold anywhere, but it laid the foundation of smart home units and research into the field.

The use of sensors and meters to pass along information using a network can be helpful for monitoring all the activities and assets remotely. Sophisticated techniques are integrated into smart homes to provide a large amount of information that is helpful for the maintenance and performance of predicted conditions. The use of persuasive technology given by [8] gives complete information about the future of smart home technology. The automation of any home for the conservation of energy can be really helpful in achieving goals, such as hydro-thermal, visual, air quality, and also plug loads usage [6].

The connectivity amongst various sensors, appliances, components and devices, with the help of a network education system, collectively creates a smart home environment. The remote access and monitoring can be done for the house with the help of these devices and observation. A centralized database-monitoring unit is responsible for recording all the observations and information about the system. On the same side, the devices are connected to provide smart energy management from remote locations, even if the owner is not present at the house [6]. The most important unit of a smart home is the network itself. The real-time exchange of information is done in such a way that the monitoring and decision-making can be done from remote locations. An advanced control system is good enough to provide these features in the smart home environment. Significant savings of 5%–15% energy consumption is predicted in the study, compiled by [9].

With the increasing charm of using a smart home, large numbers of vendors have started producing gadgets that are responsible for providing special features that are integrated with sensor-based devices. Some of the important organizations producing devices capable of providing smart home features include Apple Home, Google Home, Arduino Smart Homes, Raspberry Pi, ZigBee, etc. The integration of the services given by these companies provides smart home solutions. All the smart home devices are usually connected with the help of a network created within the home. The information that is recorded with the sensors is passed on via this network to the cloud of the organization. Once the information reaches the cloud, it is recorded and can be used for real-time monitoring and observations.

However, whenever it comes to network and information sharing, the most important part is the confidentiality and security of the information. A large number of data leaks in the previous year has revealed vulnerabilities in the sharing of information of an individual home over the network. The adaptation of technology of smart homes is the need of the hour, but the fear of information security and illegal use of the information remains a big issue. The authorization of a validated user to access the information from the cloud for remote monitoring is indeed one of the most important factors

responsible for safeguarding individual privacy. The potential loss of biometric information, such as voice samples, fingerprints, retina scans, etc., should be checked at every instance.

The Mirai Botnet Attack of October 2016 was one of a type of massive Distributed Denial of Services Attacks that sacrifice the information from millions of IOT devices. The result of such a sacrifice of information influenced millions of users whose businesses were relying on the services provided by Mirai. The organization placed the source code of Mirai Botnet on the Internet where it was sacrificed and the information from their servers was stolen and misused. With the increase in the popularity of the usage of IOT-based devices, the future of cyber security is still not clear for smart homes.

III. SECURITY AND PRIVACY IN SMART HOMES

When it comes to the security of a smart home, privacy concerns are on the highest acclivity. The success of smart homes depends on the challenges posed by the security issues. The use of electronic devices to safeguard the home is becoming popular nowadays. However, the dynamic networks created and the integration with the Internet of things devices make it challenging. Large numbers of devices are vulnerable to individual privacy, and make it very easy for the attacker with easy access, once he connects to a dynamic network [10]. Sometimes, the user of the smart home devices is responsible for yielding the space to such vulnerabilities. Incomplete knowledge about the use of such devices makes it easy for an attacker to create a threat towards individual security and policy concerns. An exponential rise seen in the frequency of attacks at such complex dynamic networks constitutes an alarming situation [11].

A. Information Storage

One of the most challenging situations is the collection of information from the devices on the vendor servers or cloud storage. The data from the smart home reaches the cloud storage of third parties; this data can be confidential, as well as critical, and may yield to a security breach. Two billion records were once sacrificed from a Chinese agency that ran an IOT-based platform [12]. The geolocation and the statistics, including the precise information of the household, may lead to burglary opportunities. There are chances for various devices rendering, based on the circumstantial availability of the data.

B. Copyright Infringement

During the initial setup of devices integrated with the smart home, terms and conditions are accepted, along with the provisioning of voice samples and sometimes-biometric recognitions [13]. This can result in a major threat towards individual privacy and concern. Even private communications may sometimes prove not to be private, since they are monitored and captured with the help of smart devices [14]. The communications are synchronized with some remote servers that are inaccessible, which in turn compromises the privacy policy of individual data.

C. Attack Probabilities

Generally, once, when a user accesses the internal network, and the devices that are active record observations, it is

probably likely that vulnerability inside the network can occur. The attacker tries to gain access over simple devices, which in turn can provide further access to a larger number of information channels throughout the network. A single vulnerability at a weaker device can be a big threat to the entire network of smart home devices. The ability to track family member habits/behavior or location after getting access to some vulnerable home devices is yet another possibility that can arise. An official watchdog was found once in 2017 to instruct parents to destroy Cayla, a doll [15]. The deciphered reason showed that one of the Bluetooth devices was compromised on security breaches by an attacker who used the doll to talk with the child playing with it.

D. Physical Security

One of the most important and key factors for the IOT-based devices is physical security. The use of low quality sensors and cheap digital circuits can be harmful and hazardous to the devices that are implanted inside the house [16]. The malfunctioning of these devices can lead to improper readings, such as for the location, and addressing of the desired information. It is also likely that these devices are weathered very easily with the normal fluctuation of temperature and moisture conditions. It is also worth noting that such cheap devices are available in the market in a variety. However, there has to be a trust with all the manufacturers and the devices that are implanted inside the home to provide smart home services.

Yet another consequence for the use of such devices is the possibility that they can provide dangerous situations of fire and destruction. Since a major portion of each device is controlled with the help of electric signals, there are chances of electric malfunction due to the use of low quality manufacturing material or doped semiconducting material that can be harmful. The handling of these gadgets requires special training, and the deployment of these devices needs an experienced workforce. However, the lack of such a workforce and an inexperienced staff can also lead to issues created inside the circuits.

E. Lack of Control and Awareness

The smart home devices and gadgets that are implanted at any house require sophisticated handling and a managing skillset. These devices are delicate and need special attention for handling. Some of the precautions that are required to handle these devices are comprised of the following facts:

- 1) The devices should be planted at a considerable height, out of the reach of children [17].
- 2) They should be located at a place where rodents and moles cannot destroy or damage the cabling or wiring of the system.
- 3) There are several devices that require restricted moisture conditions or temperature variations [18]. It should really be emphasized that the location of such devices having sensitive sensors should be away from any underlying physical conditions that do not meet the requirements.
- 4) Installation of the devices should be done precisely, and all the connections as well as the cabling must be tested and checked at proper times. The maintenance of such systems

should be done at regular intervals to assure proper working of the smart home.

5) The most important part for using such devices is proper training and awareness. All the users who are administering or using the devices should be properly dedicated for such use [19]. They must undergo proper training from an experienced organizational professional for the code of conduct and usage.

F. Integrity of Data

The most important concern for the use of smart home devices is the data. Care should be taken to safeguard the privacy of individuals. As a safety measure, the end user must know the data that is travelling in the network that is using the devices. The privacy of the data should not be compromised at any instance. It is also recommended that the end user should be a technical and technology-efficient person, to handle the information flow and its security. While there are many chances for any type of data breaching or hacker attacks, it is very important to decide which data should be flowing in the network and control the data as per the privacy concerns.

IV. RELATED WORKS – IOT IN SAUDI ARABIA

For the growth of the nation, and the Vision 2030 of Saudi Arabia, there is a big scope for adaption of smart homes in the country. To improve the quality of living and facilitate the citizens, digital transformation is going on in every sector of government as well as private organizations. Large numbers of sectors are targeting towards improvement in IOT-based devices, which includes environmental monitoring, infrastructure management, manufacturing units, transportation, medical health care and home automation. The Communication and Information Technology Commission (CITC) in the Kingdom of Saudi Arabia have given certain guidelines based on which integration of IOT devices is going ahead for various sectors in the country [20]. There are certain standards that are followed as per the guidelines generated by CITC.

However, a descriptive survey was conducted by [21], which reflected the special requirements before the application of IOT. The survey also conducted a very strong screening method along with certain analyses to identify the role of the data centers in the country for minimizing the cybercrimes and the risks. It is clear from the above section that adaptability of smart homes in a country depends entirely upon the perception of the people and various other factors. To reduce the risk and the level of cyber threats towards the information and data, a country is expected to implement powerful decisions. Proper measures for the adaptability of smart homes in any country should be taken prior to the adaption. It is clear that the seriousness of the use of IOT, in the Saudi Arabian region, can be addressed with the help of educating the users, providing security measures, managing information and privacy, handling cyber situations, and, finally, identifying the required security requirements [21].

The use of integrated technology and embedded devices for providing smart homes will open the path for cyber criminals and hackers, from which the misuse of information is more likely to happen. In contrast with the facilitation of smart

homes, cybercrimes and other non-ethical data privacy issues are more likely to occur. Internet-connected devices are going to facilitate the user on one side, but can be more dangerous at the other end [21]. As presented in one of the studies by [22], approximately 50 billion devices are connected across the globe, through which the data is travelling. This number is going to rise significantly in the near future. The IOT model for maintaining security and privacy requires specific attention to security management, including identity management as well as data ownership. There has been an inverse relationship as identified by [21] between the awareness as the former part, and readiness as the later part, in the domain of IOT and cybercrimes.

A large number of communities and cities are planning for growth and development in the kingdom of Saudi Arabia towards the proper planning and adaption of the smart city. The correlation between smart city and smart homes is really clear, and the development of the idea entirely depends upon the adaptability of smart homes in the country [23]. Large numbers of wireless sensors, in association with technology-driven routines and mechanisms, empower the uplifting of the living standards for an individual in his house. The IOT conference, organized by [24], showcases the use of IOT in Saudi Arabia. The conference also leads to case studies as well as to smart solutions that are an integral part of the Vision 2030 of the Kingdom and beyond. The adaption of IOT and its allied services are also portrayed in the NEOM City project of Saudi Arabia [23].

It has also been identified that approximately 81% of businesses in Saudi Arabia have already started using IOT and its allied technologies [25]. Big players in the software industry, such as IBM and Oracle, have also entered the race. The IOT harp for these organizations has released special devices and sophisticated software in the Middle East region for adaptability. With the use of IOT in business, it can be considered that the adaptability of smart homes is not going to take a lot of time. It is further expected that the Saudi Arabian market for IOT devices and their adaptability is going to exceed \$3 billion in subsequent years [26]. It is also expected that around 11.3 million devices will be having the connectivity towards the emerging technologies, which will be making use of IOT techniques enabled with the help of wireless sensor networks, embedded systems, persuasive computing, etc. However, along with the upcoming growth and advancement in technology, the risks are also increasing in cyber security. Technology will be more vulnerable to cybercrimes and attackers who hack valuable information and present hazards to individual privacy by data breaching.

V. RESEARCH METHODOLOGY

Under research methodology, processes for data collection will be described. This section of the research will also delve into illustrating the adopted methodology for the proposed validation. The function of the quantitative research method is

explored in this research to investigate the consumers' perception about smart home security and privacy in Saudi Arabia, encompassing usage as well as terminology. The selection and choice of this approach was informed by its capacity in determining not only opinions but also attributes of participants.

We used a survey to evaluate the respondents' concerns about the security and privacy of smart home devices in Saudi Arabia. The aim of this research is to identify security and privacy challenges facing smart homes, followed by an investigation of the users' privacy and security concerns of smart homes. In addition, this research also attempts to recognize privacy security mitigation actions that users take to protect themselves, and to know the opinions of users about those responsible for the privacy and security of smart home devices.

The questionnaire was developed into five distinct segments, comprised of demographic information, security and privacy concerns, mitigation actions, usage of smart home devices, and responsibilities. Under the demographic segment, elements such as age, job status, sex, education, and organization are covered. The second section contains smart home device types and numbers, the time period for using the devices, data types of smart homes, and the security classification level and sensitivity of data-based consumer opinion. The third part focuses on the security and privacy issues facing the smart home environment and consumers driven from the concerns that are discussed in section two of this paper.

The fourth part concerns the participants' behavior and action to mitigate risk and protect consumers and their smart home devices. Finally, the last part covers the roles and responsibilities of those involved in the implementation of smart home devices, based on the consumers' point of view. The survey exploits the Liker Scale, which is calibrated with measurements of 1 to 5, coinciding with (1) Agree (2) Disagree (3) strongly Agree (4) Strongly Disagree (5) Neutral. Different but appropriate software is used in analyzing research data and in completing the survey.

A. Data Awareness

Due to the issue of limited space in data presentation and discussion, much attention will be drawn towards relevant results that will be generated under this section with respect to data analytics. Our sample size was more than 210 participants, including 205 participants who already use smart home devices; 72% of the total number of participants use between two to five types of smart home devices at the same time, and 69% of the total number of participants have used smart home devices for more than three years, as shown in Figure 1. In terms of education level, the majority of the participants have bachelor's degrees, representing 51%, and 23.5% have a master's degree. In terms of the location of the participants, 95.4% live in Saudi Arabia, and, because they are the only target audience, other participant's results are ignored.

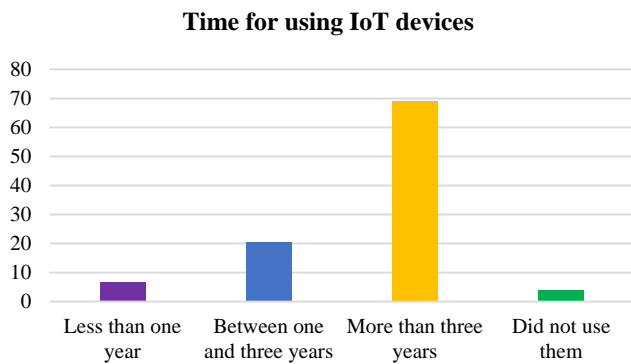


Fig. 1. Age Group based Mapping for use of IoT Devices.

In terms of demographic information, our study shows that the age group from 26 to 35 years old is the largest in our study, representing 44.9% of the total number of participants, followed by the age group from 36 to 45 years old at 35.2%. The type of living of the respondents shows that people living with their family were the largest groups in our study, by 96.9%, and 2.6% live alone and 0.5% live with friends. Despite the low number of female participants (30.6% of the total number), this study indicates that there exists a percentage in terms of consumer perception between males and females.

In terms of the purpose of using smart home devices, the majority of participants, 55.1%, use smart home devices for entertainment and relaxation purposes, while 51% of the total number of participants use smart home devices equally to receive technical assistance and to keep up-to-date with the development of technologies around the world. The result of the survey shows that most of the participants, 79.1%, are concerned about their data used in these devices. Consumers are worried about how to control and secure the data, which indicates the importance of this research, taking care of the causes and solutions, and providing adequate guidance and training.

B. Discussions

This research aims to investigate smart home security and privacy issues based on the consumers' perception in Saudi Arabia. Therefore, this section is classified into five main categories: privacy issues, security concerns, awareness and knowledge, consumer's mitigations, and stakeholder responsibilities. The main research questions to be answered in this section are these:

- 1) What are smart home users' privacy and security concerns in Saudi Arabia?
- 2) What privacy/security mitigation actions do users take in Saudi Arabia?
- 3) Who do users believe is responsible for the privacy and security of their smart home devices?

C. Privacy Issues

Proliferation of intelligent systems into the day-to-day life of people's homes stems from the rapid and robust development of the Internet of Things. But, the increasing utilization of these technologies has generated concern regarding the collection,

handling, and usage of sensitive data. Issues have been complicated by the fact that the function of privacy continues to be unexplored within the context of smart home usage. It is noteworthy that information privacy denotes the capacity of an individual to control their personal information. Achieving such a goal is increasingly becoming difficult, given the advancement of digital technologies. Different sensing technologies are explored in smart home devices and in providing services. By collecting colossal amounts of data, these sensors provide services to users by allowing these data to be processed and interpreted. However, a combination of the collected personal data and exploitation of internet-connected devices predisposes residents to emerging security and privacy risks.

Based on the survey results, 79.7% of the respondents are worried about the statutes of the privacy and security. Their data can be spread without their knowledge. There are many parties through which data can be spread, such as smart home device manufacturers and Internet service providers. Aspects of control and certainty are indicated as necessities for human beings, therefore, the absence of these elements results in suffering [27]. Our result shows that 78.8% of participants are afraid of smart home device manufacturers who can access their private data and monitor user behavior. In addition, when participants were asked about privacy violation by intrusions on their privacy, even if any third parties had obtained consent from users regarding the authority to access some consumer data, they replied as follows, and as shown in Figure 2: 47.8% strongly agreed, 30.9% agree, 11.2% neutral, 8.4% disagree, and 1.4% strongly disagree.

One reason for the growing consumer concern is the weakness of their knowledge about data collected and when and how it is gathering. Because of low levels of public awareness, functionality continues to be the primary focus of many users. Therefore, they fail to interrogate how these services are being provided [28]. The second reason may be due to the weakness of the standards and policies of protecting user privacy that need to be fulfilled by smart home device manufacturers, or which perhaps exist but are not present in the correct and understandable manner. Regulations and standards should be implemented in the products and services provided by companies, whether they are local or global, such as by the General Data Protection Regulation (GDPR). Privacy policy must be provided by vendors of smart home devices, as such policy is regarded as an internal statement designed to guide an entity or organization in handling and managing personal information. As a matter of fact, it is targeted at the recipients of personal information. Employees are instructed on the correct approaches of collecting and using data as sanctioned by the privacy policy on the collection and the use of the data.

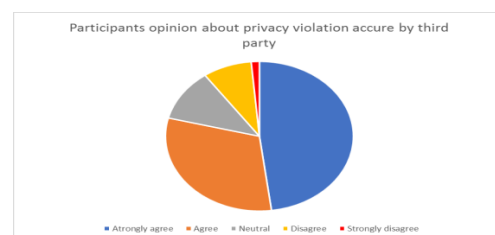


Fig. 2. Trust Percentage on Privacy Violation by ISP.

The result shows that 41.3% of the respondents trust the Internet service provider, while 27.6% express concerns about their data. Government monitoring of the performance of Internet service providers to ensure quality may increase user trust to use smart home devices. In Saudi Arabia, the regulatory framework for Cloud Computing is established to protect user data, which must remain stored in cloud services within the Kingdom of Saudi Arabia [29]. In general, privacy is one of the main concerns of people and of whether users are concerned over their inability to control personal and private information. Such concerns are measured to make users aware of the risk of invasion, thus minimizing privacy issues. Applying this to the context of smart home devices will make users shy away from adopting these technologies, out of fear of possible breaches to their privacy.

D. Security Concerns

The fact that a security technology is easier to use might hint to the fact that it is easier to be intercepted or compromised by criminal minds. Each feature of a device or service is laced with a possible risk for attack as well as potential for failure. Within the smart home environments, it is difficult to design perfect watertight security. This is because of the prevailing heterogeneous ecosystem defined by a plethora of devices as well as services. Matters are complicated further by the fact that such systems, apart from having limited security, are also affected by weak capacities in terms of segments such as battery and CPU. Consequently, the provided services depend on remote infrastructures including cloud storage and analytics. It is estimated that 80% of IoT devices are open to a myriad of attacks [30]. As a matter of fact, linking traditionally 'stand-alone' smart devices, such as appliances, lights, and locks, poses innumerable cyber security risks. Some of the commonplace cyber security attacks and threats on Smart Home devices comprise: Man-in-the-middle, Device hijacking, Permanent Denial of Service (PDoS), Data and identity theft, and Distributed Denial of Service (DDoS).

According to our survey, most of the participant, 80.7%, are anxious of security vulnerabilities that exist or are likely to exist in smart home devices, including attacks on vendors. Security vulnerability refers to a vulnerability which can be abused by a threat actor, including attackers, by crossing privilege boundaries with the aim of performing unauthorized actions within the context of a computer system. In order to exploit an area of weakness, an attacker must leverage at least one applicable technique or tool that is connected to the identified weakness in the system. In this regard, vulnerabilities are referred to as the attack surface. The seriousness of the vulnerabilities is especially pronounced when they are discovered but have not yet been updated or fixed, such as in the case of a zero-day attack, which is a point of exploitation on the computer software that remains hidden from those who are supposed to alleviate existing vulnerabilities, such as the vendor of the vulnerable software. Unless these points of weakness are minimized, hackers can take advantage of them to disorient and compromise computer data, programs, networks, or additional computers.

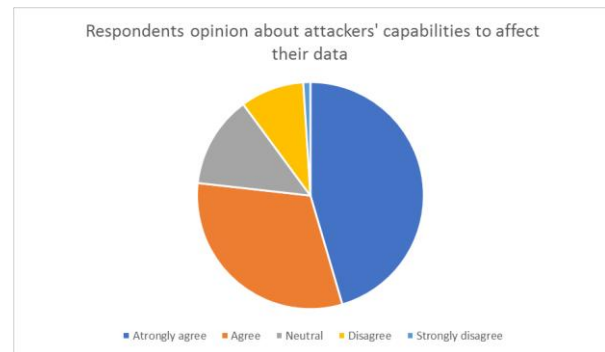


Fig. 3. Percentage of Opinion for Vulnerability towards Data.

When respondents were asked about believing that attackers have capabilities that enable them to access user data illegally and negatively influence it, the respondents replied as follows, and as shown in Figure 3: 45% strongly agreed, 31.4% agree, 12.6% neutral, 9.3% disagree, and 1.4% strongly disagree. The high percentages may reflect feelings of insecurity, especially since much of their data—such as medical, financial, and other data—can be accessed through smart home devices. The participants feel that the data they are most scared of being leaked are financial data, by 75.1%, then family data by 68.5%, then video data by 66.7%, then national ID number by 53.1%. On the other hand, smart home device performance monitoring data received the lowest level of importance, based on the user's perspective, with an 8.5% rate.

Some vendors may use what is known as off-the-shelf software or commercial off-the-shelf, which are products that are highlighted as packaged solutions with attributes that are designed to meet the needs of the purchasing organization, as opposed to bespoke, or custom-made, solutions. The disaster is that the breach of such software or platform could affect all companies and users who are working on it, as happened with the Solar Winds company in the United States recently, and this is what most of the participants in the questionnaire believe, with a percentage of 64.3%. It is evidently clear that cybercriminals are on a hunt for IoT-related points of weakness that are especially lacing new devices, and this communicates the need for adequate security to be implemented at the design phase and along the different stages, up to the deployment phase of the device. What this may hint at is the fact that vulnerability management associated with IoT devices has an important function in reducing attack openings.

In general, security is one of the biggest challenges facing smart home devices. Most individuals watch favorite programs and film Smart TV, and even connect baby monitors to a home network, but, in many cases, they do this in total disregard of their devices' security. It should be noted that one single mistake can give way to a hacker. While on the network, people can gain access and connection to other devices. This means they can abuse the accessed personal data information of voice recording and video, including streaming and storage. Security is an important part that needs to be addressed by all parties, as discussed in the next section, on responsibility.

E. Awareness and Knowledge

Awareness among consumers concerning privacy and cybersecurity issues is underlined as a crucial factor of organizations' posture on cybersecurity. It is understood that an adequately informed consumer regarding key issues related to cybersecurity will stand a better chance of defending and guarding against social engineering as well as other attacks. In order to appropriately develop IoT resources that will serve the communities better, the creation of a higher degree of awareness with regard to cybersecurity is required. This should go hand-in-hand with the training of users. This is the best approach, positioning consumers as active players and as a defensive layer within the structure of organization's security. Achieving this end requires initiatives aimed at educating and instructing the public with the primary aim of diminishing risks related to the IoT environment. The first step is to identify the status of security awareness regarding IOT vulnerabilities, and then design a program that includes best practices for increasing the level of awareness. In Australia, the government published and released.

The Code of Practice in August 2020 with the aim of strengthening the security of the Internet of Things for users. Both the Voluntary Code of Practice and The Code of Practice-Securing the Internet of Things (IoT) for consumers are established on 13 principles. However, our research found that 39.4% of respondents believe that cyberattacks are a fact and not an exaggeration, while 38% think it is not true, and 22% of respondents were neutral about this issue. This indicates the presence of partial awareness among the participating group, some of whom may not have previously been exposed to a cyberattack. In addition to the above, three-quarters of the respondents feel that they have sensitive data that needs protection. Reading instructions and training on the use of smart home devices is one of the most important practices that indicate increased awareness of the user, as the results have shown that 69% of the participants believe that reading the instructions and asking about them and how to use them safely is a necessity. In addition, 46% of respondents believe that insecure devices that do not give the user the authority to control and manage access to data in these devices should not be purchased. On the other hand, 50.7% of the participants follow news and technical posters related to the security of smart home devices, while 15% do not show any interest, as shown in Figure 4.

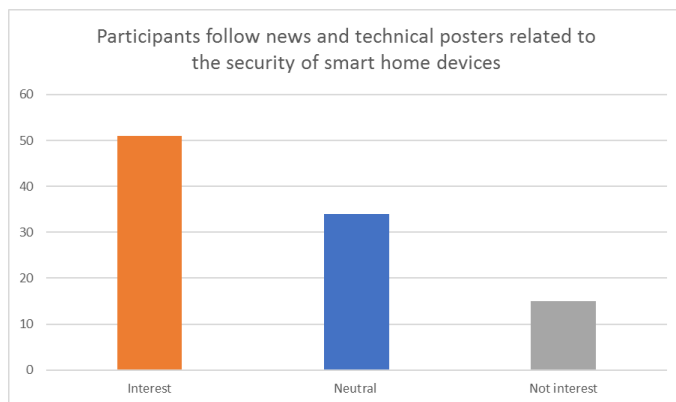


Fig. 4. Smart Home Security Adaption by Users.

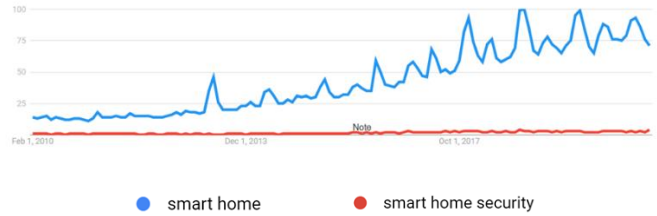


Fig. 5. Trends in 'Smart Home' & 'Smart Home Security'.

In Figure 5, comparison is made between the search terms 'smart home security' and 'smart home'. The aim here is to determine if users interested in adopting home automation technologies are equally keen on ensuring that their devices are secured. It should be noted, however, that this should not be misconstrued to mean that, in so doing, the research is aiming at a definitive statistic of trend. As a matter of fact, this is merely an approach that offers more insight into the content of the literature review. Assessments indicate that, while a focus on smart homes has risen significantly over the past 11 years, the attention given to the security of devices is still insignificant. This reality raises concern, since not all users are technologically oriented. Therefore, it is likely that these users are unaware of pertinent privacy and security concerns facing them.

F. Consumers Mitigation

Process and policies adopted with the aim of minimizing data breaches and security incidents alongside the diminishing extent to which damage may occur is referred to as cyber security threat mitigation. But the pertinent question remains: what is the role of a consumer in mitigating data breach risk and thwarting hackers? There are various options designed at achieving mitigation function. These can range from simple low-level actions performed at a personal level, to organization-wide business strategy changes. It should be underlined that a number of simple practices and rules, when properly followed, can position individuals tasked with sensitive data at the required vantage point. This goes a long way in diminishing and preventing the degree of exposure to cybersecurity risks. Five categories of simple function designed for consumer mitigation exist: identify, protect, detect, respond, and recover.

When asked about the practices they take to secure their smart home devices, the practice of using a strong password to prevent unauthorized access ranked first, at 88.2%. The second practice is keeping data about devices safe and not sharing it with anyone outside the house, at 84.9%. The next practices are updating the IoT devices when the update is published and asked by device makers, as well as minimizing permission given to lower levels to access data, at 83.9% equally for each practice. The next practice is securing the Wi-Fi network inside the house and encrypting the data, at 82.1%. After that, making sure that all settings are properly and securely well configured, and working on the settings carefully with as few mistakes as possible, at 80.7%. Finally, asking and only buying the devices that consider security and privacy, ranked last, at 74.1%.

It should be underlined that every individual faces similar threats against organized attack, thus, every individual is

expected to protect all data and anticipate the worst possible attack scenario. Individuals who fail to regularly secure their software within the context of an IoT environment are likely to be attacked by sophisticated hackers. Previous practices would achieve cybersecurity goals in smart home devices.

There are three key primary goals of information security: (1) preventing the loss of availability, (2) the loss of integrity, and (3) the loss of confidentiality for data and systems. Most security controls and practices are designed to eliminate losses associated with each of the highlighted concerns. The aforementioned elements together form the AIC security triad. This represents the initials for availability, integrity, and confidentiality.

G. Stakeholders Responsibility

Given the fact that the risk factor has been clearly identified, has this helped in enlightening who should carry the responsibility of IoT security? Because of the many different participants in the operation and maintenance of security devices, there is a significant level of uncertainty. Guided by our research survey, 89.2% of the total respondents indicated that the one single body to bear this responsibility is the government, 74.6% indicated that consumers who are users of the devices are to carry this responsibility, and 74.1% placed the burden of responsibility on the manufacturers, as illustrated in Figure 6. All of these responses contain some degree of truth from the fact that each of these entities must bear responsibility, insofar as achieving a comprehensive IoT security is concerned. However, it is understandable that most respondents in the survey hold the position that governments should bear responsibility, being the main stakeholder of IoT security. Attaining this end requires that governments develop and implement policies and legislations aimed at monitoring and controlling the IoT market. This will serve to ensure compliance regarding the stipulated policies. In fact, the Ministry of Communication and Information Technology should use best practices and theory in IT as a way of strengthening the overall cybersecurity within the context of IoT.

Providers of IoT-enabled devices, such as manufacturers who are part of the security system, must effectively communicate and educate end users or integrators of any possible risk. By illustrating a commitment to protect users of their equipment, manufacturers are viewed as both understanding and trustworthy in the eyes of users. This can be attained by providing the necessary education to users. Another key and vital step that manufacturers can explore is encryption between devices, as a way of reinforcing protection within the IoT system. It is instructive that best protection with regard to data protection is developed whenever consumers use devices that are connected to a network. Some of the approaches of attaining this end are by disabling default credentials, practicing the safe sharing of sensitive information, use of proper and adequate password etiquette alongside the instinct to avoid any questionable requests or activities. Consequently, one of the best approaches of diminishing any misunderstanding regarding IoT security responsibilities is by roping in every contributor of the IoT enabled devices. Although concerns and fears regarding IoT are rife, a guarded and secure system can be attained by bringing together the manufacturer, the organization, and user.

H. Implication

Several security as well as privacy concerns have emerged because of the rapid and widespread adoption of technology. Moreover, the interconnectivity of the various different appliances has served to increase such issues. This comes at the backdrop of the fact that many users are oblivious to the dangers and risks of connecting their devices to their home network. It should be underlined that data generated and collected from smart devices can expose colossal volumes of personal information, for instance, likes, dislikes, and daily routines, thus revealing a lot about the user. Governments must take care of policies and legislation that protect users and preserve their privacy, while indicating the penalties for violating these regulations along with legislation to limit their abuse.

The United Kingdom also took the initiative of developing and publishing The Code of Practice for Consumer IoT Security, founded on 13 outcome-focused guidelines encompassing what is regarded as the generally acceptable tenets of good practice within the IoT security domain. In the same vein, the Canadian government has published a common understanding around the Internet of Things. This is highlighted as important in safeguarding the privacy of consumers within the IoT environment, as highlighted by the Internet of Things Security for Small and Medium Organizations.

However, as governments adopt such measures it should be acknowledged that individual users are expected to exploit best practices in addressing cyber perils as well as improving their levels of awareness by virtue of education.

Rating the security and privacy for each device that they purchased in the market is important to make other people aware and to notify vendors to pay attention to these issues. Vendors must provide transparency, including what data is being collected, in order to give the consumer the ability to control and manage their data with different options as well as to update the device to protect it when it faces an attack.

It is expected that organizations balance their need for connectivity and efficiency embedded in IoT technologies with risks and threats emanating from this connectivity. This is particularly the case given the prevailing absence of security-oriented design and development of these products.

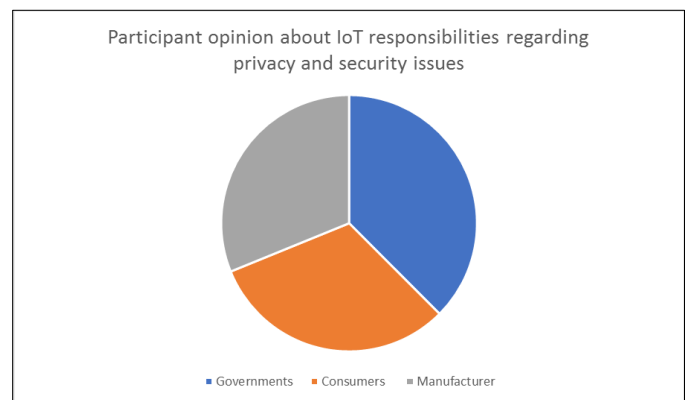


Fig. 6. Privacy and Security Responsibility Chart.

In order to prevent cybersecurity attacks within the context of the IoT environment, there is need to adopt a number of useful technical solutions by both consumers and vendors. This is informed by the reality that consumers are expected to create their layer of defense. One of these approaches is the use of multiple routers in order to be able to set up optional networks. Similarly, a router possesses the ability to separate different computing devices from IoT devices by directing them towards different numerous available networks. This makes it difficult for cyber criminals to attack effectively, because an attack on one device does not affect other devices on other different networks. To attain this end, consumers are required to check through reviews in order to get recommendations. Consequently, they are expected to perform research on security capabilities, buy or source their IoT devices from vendors and manufacturers with clear track records, and set automatic updates on their devices active for any available update.

VI. CONCLUSION

This research focuses mainly on the identification of all the security challenges that can arise because of the use of a large number of IOT devices connected to provide a smart home facility in Saudi Arabia. The beginning stage of the research focuses on the identification of the factors responsible for security concerns in smart homes. The research also tries to identify and provide various actions that can be taken in accordance with the problems associated with the protection of individual privacy and security against threats that might arise in the country. A realistic quantitative methodology is designed to identify the variety of people who are affected due to such security threats. However, it is really clear that, wherever data exists, there are probability and chances for concerns towards data stealing, breaching, hazards, etc. Therefore, the entire research is focused mainly on a large number of factors that are responsible for any security or privacy threat. Towards the conclusion of the model adapted in this research, benchmarking results are expected to be obtained and analyzed for the minimization of problems associated with security and privacy threats in Saudi Arabia.

REFERENCES

- [1] Weiser, M. Ubiquitous computing. in ACM Conference on Computer Science. 1994.
- [2] Khan, Y. 5 Essential Components of an IoT Ecosystem. [English] 2020 21 April 2020 [cited 2020 21 April]; Available from: <https://learn.g2.com/iot-ecosystem>.
- [3] Sato, H., et al. Establishing trust in the emerging era of IoT. in 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE). 2016. IEEE.
- [4] Program, Y.E.-G. The National Strategy for Digital Transformation in Saudi Arabia. Digital Transformation [English and Arabic] 2020 24/11/2020 2020]; 1:[GOV.SA]. Available from: https://www.my.gov.sa/wps/portal/snp/aboutksa/digitaltransformation!/ut/p/z0/04_Sj9CPyksy0xPLMnMz0vMAfIjo8zivQIsTAWdDQz9LUxNnA0Cg11DXEydAowCHQ31g1Pz9AuyHRUB1eTRhg!/.
- [5] Aleisa, N. and K. Renaud, Yes, I know this IoT device might invade my privacy, but I love it anyway! A study of Saudi Arabian perceptions. 2017.

- [6] Fabi, V., G. Spigliantini, and S.P.J.E.P. Corgnati, Insights on smart home concept and occupants' interaction with building controls. 2017. 111: p. 759-769.
- [7] © Maevi Sdn Bhd | 2017 -2020, A.R.R. The History Of Smart Homes. The Beginning of Home Automation 2020 [cited 2020; Available from: <https://maevi.my/the-history-of-smart-homes/#:~:text=In%201966%20%E2%80%93201967%20%E2%80%9320ECHO%20IV,turn%20appliances%20on%20and%20off>.
- [8] Emeakaroha, A., et al., A persuasive feedback support system for energy conservation and carbon emission reduction in campus residential buildings. 2014. 82: p. 719-732.
- [9] Nilsson, A., et al., Smart homes, home energy management systems and real-time feedback: Lessons for influencing household energy consumption from a Swedish field study. 2018. 179: p. 15-25.
- [10] Heartfield, R., et al., A taxonomy of cyber-physical threats and impact in the smart home. 2018. 78: p. 398-428.
- [11] Ferrag, M.A., et al. Privacy-preserving schemes for fog-based iot applications: Threat models, solutions, and challenges. in 2018 International Conference on Smart Communications in Network Technologies (SaCoNeT). 2018. IEEE.
- [12] Wang, Z. Personal information security risks and legal prevention from the perspective of network security. in The International Conference on Cyber Security Intelligence and Analytics. 2020. Springer.
- [13] Lopatovska, I.J.U.J.o.L. and I. Science, Overview of the Intelligent Personal Assistants. 2019. 3: p. 72.
- [14] Hall, F., et al., Smart Homes: Security Challenges and Privacy Concerns. 2020.
- [15] Oltermann, P.J.T.G., German parents told to destroy doll that can spy on children. 2017.
- [16] Atlam, H.F. and G.B. Wills, IoT security, privacy, safety and ethics, in Digital Twin Technologies and Smart Cities. 2020, Springer. p. 123-149.
- [17] Sivaraman, V., et al., Smart IoT devices in the home: Security and privacy implications. 2018. 37(2): p. 71-79.
- [18] Abraham, E.A.J.A. and I. Research, The Challenges Of Security In Internet of Things (IoT). 2019. 6(3): p. 165.
- [19] Lim, H.-K., et al., Federated reinforcement learning for training control policies on multiple IoT devices. 2020. 20(5): p. 1359.
- [20] CITC, K. Quality Of Service Indicators 2020. 2020; Available from: <https://www.citc.gov.sa/en/Pages/default.aspx>.
- [21] Alanazi, M.H. and B. Soh, Investigating cyber readiness for IoT adoption in Saudi Arabia. 2020.
- [22] Plachkinova, M. and P.J.I.S.F. Menard, An Examination of Gain-and Loss-Framed Messaging on Smart Home Security Training Programs. 2019: p. 1-22.
- [23] Alam, T., et al., Big Data for Smart Cities: A Case Study of NEOM City, Saudi Arabia, in Smart Cities: A Data Analytics Perspective. 2021, Springer. p. 215-230.
- [24] IOT, S. Annual Saudi IOT Conference. 2018 [cited 2018; Available from: <https://saudiidot.com/iot-conference/>.
- [25] SaudiGazette. Business In Saudi Arabia. 2020; Available from: <https://saudigazette.com.sa/article/590331>.
- [26] Al-Khattaf, D.-I. Saudi Arabia's IoT Market to Exceed \$3 Bln by 2023. 2020; Available from: <https://english.aawsat.com/home/article/2659731/saudi-arabia%E2%80%99s-iot-market-exceed-3-bln-2023>.
- [27] Siegel, D.J.C.M., available at <http://changingminds.org/explanations/needs/control.htm>, The need for a sense of control. 2008.
- [28] Weinstein, M.J.H.P., What your FitBit doesn't want you to know. 2015.
- [29] CITC, K. Regulatory materials on cloud computing;. 2021 [cited 2021; Available from: <https://www.citc.gov.sa/en/RulesandSystems/RegulatoryDocuments/Pages/CCRF.aspx>.
- [30] Cekerevac, Z., et al., Internet of things and the man-in-the-middle attacks—security and economic risks. 2017. 5(2): p. 15-25.

Smart Company System using Hybrid Nomenclature of Neural Network

Mbida Mohamed¹, Ezzati Abdellah²

Department of Emerging Technologies Laboratory (LAVETE)
Faculty of Sciences and Technology Hassan 1st University
Settat, Morocco

Abstract—Physically, to manage the data related to the products, CRM, suppliers and Administration warehouse of the company makes us use a lot of human resources, and a time which deals with this, consequently the error rate increases and sometimes everything goes out of control, however, this work designed an intelligent overall management system (an intelligent neural network) which completes and up-date the product management network that presented in one of the previous articles. This new version assembles the three modules, in an order to automate tasks in the real time.

Keywords—Neural; network; intelligent; CRM; company; warehouse; real time

I. INTRODUCTION

A. Standard Logistics Inventory Administration

These days, when talking about stock administration, it's truly about keeping and following an amount of merchandise in a store. It is an action that separates into: the board of developments passage and leaves products; recharging the executives; lastly a related assignment the administration of the article documents. Contingent upon the association of the organization, these errands can be doled out to an individual or two diverse staff profiles.

The vendor for the execution of the physical developments of passage and exit;

- Supply Management for stock following and recharging
- Agent gathering data about every item by a gadget standardized tag committed to peruse the bare-code scanner.

Notwithstanding, most industry divisions use standardized tags from numerous points of a view.

Standardized identifications have altered the creation, preparing and observing of items in the food, bundling, retail, clinical, a drug, gadgets, a car, mechanical segments and air transportation.

Bare code Scanner can be found on all electronic and mass-market items, from the battery of your cell phone to the case containing your new sport shoes. The utilization of 1-D and 2-D codes decreases overhead expenses via mechanizing and disentangling a gracefully chain the executives, stock, choice, and the buying cycle. The modern segment has

additionally embraced standardized identifications for reasons of the security and duty.

Lately, various nations have started to require makers of clinical and drug items to put precisely comprehensible codes on all bundling, including singular medication bundles. On the off chance that a flawed item is conveyed to a drug store, the programmed following of each bundle will quicken the review of items while guaranteeing the accessibility of the value control information all through the flexibly chain.

B. Basic Customer Relationship Management

Most companies use macro dashboards to manage customer data, which can be tedious to enter and update when there are many customers, or they are using phone calls to prospect the customer's opinion in order to develop the quality of service, this can generate errors which can be costly to the company, which implies to find a solution to automate in real-time all these transactions with the customers without fail.

C. Standard Supplier Service

Each company that manufactures a finished product, needs raw or semi-finished materials, so a continuous relationship with suppliers is required by email or a phone, to have traceability it needs to save invoice information suppliers, then an intelligent and flexible solution is necessary so as not to weigh down the task for a person who will do this again a hundred or more times.

D. Classical Neural Networks

By relationship with organic neurons a network of artificial neurons must have the option to learn and recreate "smart" thoughts in an artificial mechanism (every inter-neuronal network connection will have the option to adjust and develop as learning advances).

The proper neuron is a model that is portrayed by an inside state $s \in S$, input signals $X_1 \dots X_p$ and the activation function (Formula 1)

$$S = h(X_1 \dots X_2) = \varphi(\alpha_0 + X_j) = \varphi(\alpha_0 + \alpha' X) \quad (1)$$

The activation function [1] plays out a change of a relative mix input signals, α_0 , steady term, being known as the predisposition of the neuron. This relative mix is dictated by a weight vector.

$[-\alpha \ 0 \dots \ \alpha \ p]$ related with every neuron and whose qualities are assessed in the learning stage.

They comprise the memory or disseminated information on the network. The various versions of neurons are recognized by the idea of their activation function. The principal types are:

- Linear g is the identity function,
- Threshold $g(x) = 1$ $[0; +\infty [(x)$,
- Sigmoid $g(x) = 1 / (1 + e^{-x})$,
- ReLU $g(x) = \max(0, x)$ (rectified linear unit),
- Radial $g(x) = \sqrt{(1/2) \pi} e^{-x^2/2}$,
- Stochastic $g(x) = 1$ with probability $1 / (1 + e^{-x/H})$,
- otherwise 0 (H acts as a temperature in a simulated annealing Pseudo Code).

This article, describe a hybrid intelligent architecture made-up that joins neuron networks as a learning system on Item Data, CRM and Suppliers interaction, in order to automate the general system of the company.

II. RELATED WORKS

Neural Network (NN) innovation has been effectively applied in numerous business territories, and a few works have been done, we quote the most significant

A. Estimating Model of Supply Chain Administration based on Neural Network

Pr HongJing Liu continue the utilizations of neural organization innovation [2] in gracefully chain the executives, which contain three spaces: streamlining, estimating and choice help. Nonetheless, the Back Propagation (BP) the NN is applied to gauge the interest of the bicycle in a definite area, the estimating result show that BP neural network has more noteworthy guaging exactness than that of customary determining model.

B. Artificial Neural Network for Transportation Infrastructure Systems

Artificial Neural Networks (ANNs) depict as the general interconnection of the frameworks along with numeric weighting that can be tuned dependent on experience, framework Inputs, Processing and Outputs. Additionally, the genuine bit of leeway of ANNs is the capacity to clarify complex framework issues, for example, one which are found inside the Transportation Infrastructure System's [2]. ANNs for Transportation Infrastructure System must consolidate framework designing strategies that will be economical for future years and kept up at satisfactory levels. Appropriately, Though Pr Koorosh Gharehbaghi present the idea of ANNs and its center capacities for the advancement of Transportation Infrastructure Systems specifically the support measures.

C. Predicting Logistics Delivery Demand with Deep Neural Networks

Conveyance time impacts the logistics route coordination's, contingent upon the necessities of the place and quantity. An effective forecast of conveyance request would help the organization of logistics model. The data on

conveyance request are time-reliance and space-relationship. Displaying the multidimensional grouping or building the expectation dependent on it would be a calculation expending work. Anyway this examination depends on profound figuring out how to propose an effective strategy to foresee conveyance request. With the reenactment study, the expectation performance [3,4] of the proposed technique is satisfactory. This is helpful for the further investigation of coordinations logistic choices making.

D. Social CRM and Suppliers using Web Mining

Conventional CRM (Customer RelationshipThe board) contains three modules, Marketing, Sales, and Backing, which depend on the client relationship and profiling data. While the data contained in those three modules is contribution by administrator, N. Karna and others [5] accumulate considerably more data from the Internet. can discover connection between clients and discover their profile from the Internet. This data can be utilized to improve and coordinate the CRM to perform better in supporting the business destinations. Social event data from the Internet implies that need Information Recovery and Information Extraction that include numerous sources from Internet, for example, web-based media, net blog, and news. This research gives the model of information mining usage in customary CRM to become social CRM. This exploration contributes for CRM upgrade where client driven application gets robotized.

III. SMART COMPANY CONTROL SYSTEM BASED ON HYBRID NN NOMENCLATURE

Each company has three essential modules (CRM, Suppliers, Administration inventory) to properly manage the manufacturing processes, however this work implement an intelligent system that automates tasks without human intervention by a neural network that saves and continuously learns, that will be detailed later in this article, which will follow module by module (Fig. 1).

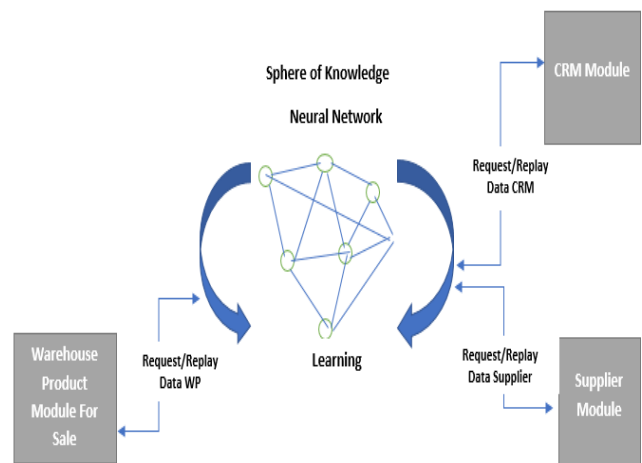


Fig. 1. Smart System Company in Interaction with the Four Modules.

A. Smart Administration Warehouse of Product Module for Sale

1) Nearest Neighbor data classic inventory administration: Nowadays Tendency is the mobility, if the data capacity is away from the PC, and the items are profound, it can't carry the items to the PC: it is important to works distantly. Some might want to deal with their stock with a phone, yet they are not "made for". There are additionally tablets or PDAs, however practically speaking you have your hands involved. That is the reason they utilize moderately basic bare-code readers, whose solitary capacity is to peruse standardized identifications and enter amounts. These readers are more powerful contrasted with tablets or PDAs, and particularly more affordable (Fig. 2).

The work is done in two phases:

- Scan items and quantities in shelves;
- Loads information saved on the PC.

This method of activity called "cluster" is sheltered and dodges dull control on a little screen. Anyway the filtering of the items just as the taking care of the data on PC, requires strategic operators to do these undertakings which can create errors during the assortment or stacking of the information (Forgets of checking/capture of the items ...).In expansion another detriment can be in the loss of time in the two stages to have an information base took care of, on account of an update of the capacity stores need to experience similar advances, which produces some other time lost. During the connectivity of the hand shower with the PC, there might be network issues (driver, connector, coding...) Or the inside memory of the code bar reader might be depleted whenever since its stockpiling is restricted to 8MB. For this the current work has built up an insightful stock administration system to lessen the pace of mistakes that can happen in this cycle of information extraction and obtaining, and furthermore to have ongoing information accessibility.

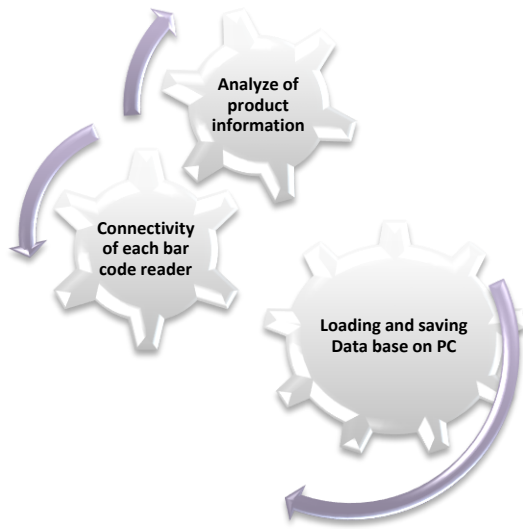


Fig. 2. Classical Inventory Administration.

2) Hybrid nomenclature implementation for warehouse product module for sale: Today, the traditional stock administration causes many losses of time and human resources just as issues with information securing. To advance this administration cycle, this requires the improvement of an incredible powerful hybrid framework dependent on Cognex technology of vision 1.2-dimensional scanner's, with sensors set up in the distribution center and an intelligent neural network with a bar-code identification, which permits the reinforcement and learning of an information in these circles of data. The execution of this design requires a section through three stages introduced in the clear figure above (Fig. 3), in what follows this article will clarify each part and its fuction [5].

3) Collection of information produced by 1D / 2D vision technology sensors: When gathering item informations, this task need the vision sensors technology, which are such an advanced camera set before each item that snaps a photo of the code. A microprocessor running an special picture preparing programming distinguishes and interprets the code before moving the acquired information to the neural network (Fig. 4) [5]. One of the fundamental elements of decision of a picture sensor, or camera, is its goal picture, in other words the quantity of pixels that make up every photo.

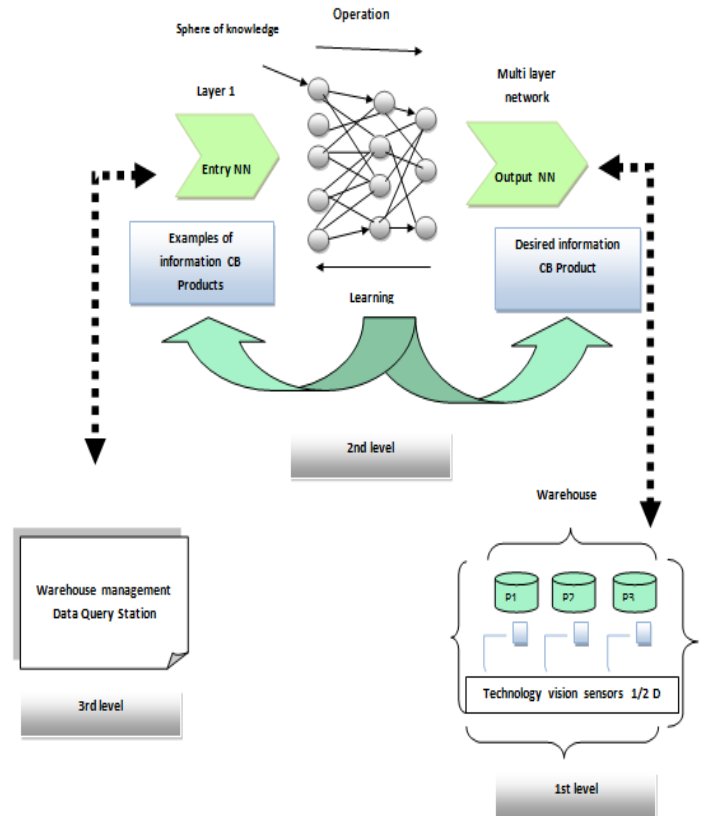


Fig. 3. Hybrid Process for Smart Inventory Product Administration for Sale.

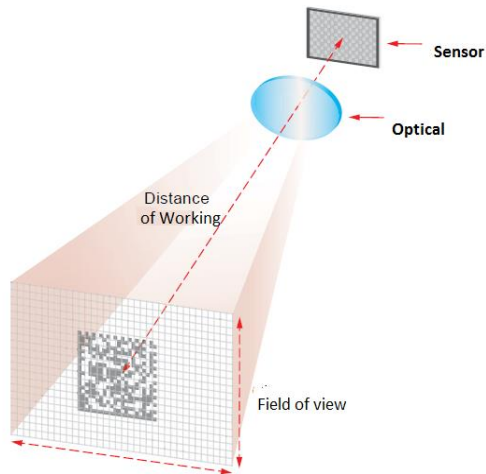


Fig. 4. Pixel Grid of the PPM Value on a Data-Matrix Code using a Vision Technology Reader.

According to a vision technology reader, the pixels resolution [6] per module (PPM). PPM is the quantity of pixels comparing to a phone or module code and ensures the camera has enough resolution to peruse this code. This worth is determined by separating the resolution of the camera one way (for instance, 752 pixels for a standard resolution readers) by the quantity of code modules, at that point doing likewise count the other way, before to increase the two numbers got. This may appear to be convoluted, yet the design applications or picture preparing programming of the mechanical vision technology code readers can rapidly compute the PPM value (Four in the model above). In expansion, new microprocessors and CMOS computerized sensors additionally empower vision-based readers to be close to as quick as the best laser scanners. These advances are notwithstanding the conventional advantages of vision-based drives: no moving parts, longer lasers, capacity to peruse damaged or omnidirectional codes, and the capacity to record pictures for review and following purposes or then again to screen code stamping frameworks[7].

Nonetheless, after the gathering of the item data as indicated by the EAN 13 Standard [8], the information is sent back to the neuron network with Wifi, since this vision innovation offers a scope of mechanical correspondence conventions including Ethernet, USB, RS-232, advanced I/O, Ethernet/IP, PROFINET and Modbus TCP/IP.

4) *Getting and learning data by NN of administration inventory product module for sale* : After the period of data extraction by vision technology sensors, they send these information to the neuron network that was planned with the Java language, to begin gathering and learning out about every item and remember it , in case if it exists in the underlying data base (.dat). The NN has been planned in Java language that permits to know and devise the scanner bar-code picture (1D/2D) as a progression of digits in four sections, so as to perceive each part as indicated by the contry , this manufacturer of this one, item producer and the control key that permits to approve the bare code. Anyway, learning

process product data in the accompanying Java code Part (Pseudo Code 1):

```
Pseudo Code 1: Getting and learning process for NN  
  
Input:  
Int N: Number of sensors of technology vision barre code  
List of series bar code received from RCSF: LSCBWN  
List of series bar code received from fichier des exemples ( ExRN.dat): LSCBF  
  
Output: Informations about products  
For each Lwsni ∈ LSCBWN[N] do  
If the Country of Lwsni > 300 && the Country of Lwsni < 379  
Sendto station (" the product is from France ") ;End;  
  
⋮  
  
Lwsni ← Lwsni+1;end;  
For each Lwsni ∈ LSCBWN[N] do  
For each LFi ∈ LSBF do  
If the manufacturing contry of ( Lwsni) = the manufacturing contry of ( LFi)  
Sendto station (the manufacturing contry of ( Lwsni)) ;  
Lwsni ← Lwsni+1;  
else LFi ← LFi+1;  
else Save into file of exemples (the manufacturing contry of ( Lwsni));  
Lwsni ← Lwsni+1; end;  
For each Lwsni ∈ LSCBWN[N] do  
For each LFi ∈ LSBF do  
If the Product of manufacturing of ( Lwsni) = the Product of manufacturing ( LFi)  
Sendto station (the Product of manufacturing of ( Lwsni)) ;  
Lwsni ← Lwsni+1;  
else LFi ← LFi+1;  
else Save into file of exemples (Product of manufacturing of ( Lwsni));  
Lwsni ← Lwsni+1; end;  
End;  
  
if ((series cb[i].ProductFab)==((seriescbF().ProductFab))  
{  
System.our.println("The manufactured product is xxxx ");  
}  
if ((series cb[i].Validkey)==((seriescbF().validkey))  
System.our.println("the key is valid ");  
}  
.....  
End;
```

B. Smart CRM Module

According to the general literature, CRM or client relationship management is a methodology for dealing with the connections and cooperation of an organization with its clients or expected clients, which requires flexibility and reduced data processing time to guarantee customer satisfaction, for this reason a CRM module has been designed which is based on an intelligent neural network by learning the customer character and offering him the appropriate product services at its choice, with an inspection of the comments ,and remarks delivered by the latter to ensure a continuous and real-time improvement of the products.

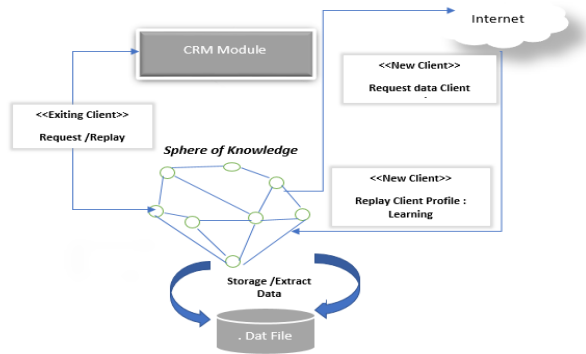


Fig. 5. CRM Module Process Managed by the Smart Company System NN.

Pseudo Code 2: Getting and learning Smart CRM process

Input:
Int L number of Existing Client
 Existing Client profile: **ECP**
 List of Product Client Profile: **LPCP**
 New Client Profile Choices product: **NCP**
 Neural Network: **NN**

Output: Informations about Client profile products choice ICP
Request ECPx
For each ECPi ∈ ECP [L] do
If ECPi =ECPx
Listing to Client LPCP
Else
Collecting and analyzing NCP From Internet
Saving Data NCP in (.dat) File
Endif; End ;

According to the descriptive figure (Fig. 5), the CRM module consults the base of the neural network in case of a customer already registered and displays the choices of these purchases according to his customer character, in the case of a new customer the neural network solicits the internet to collect profile data and analyze its product choice behavior and save it in the database linked to the NN (Pseudo Code 2).

C. Smart Supplier Module

Every productive company needs suppliers, and this must leave continuous contact with them, for this reason it was thought to automate the monthly demand for raw material products by sending invoices to these suppliers by email according to a configured period of time, and the human intervene only in the case of an update of quantity or stop of cooperation with the suppliers, this facilitates the task to the company in an organized way, the Pseudo Code 3 and the Fig. 6 below illustrate this procedure.

Pseudo Code 3: Getting and learning Smart Supplier Process

Input:
Int TER: Time for Emailing the request
Int Q: Quantity of raw material
List of Spplier Profile: LS
Int N: Number of Suppliers

Output: Send information to suppliers
For each LSi ∈ LS [N] do
If Ti = TER
Send (Request, Qi)
Endif;

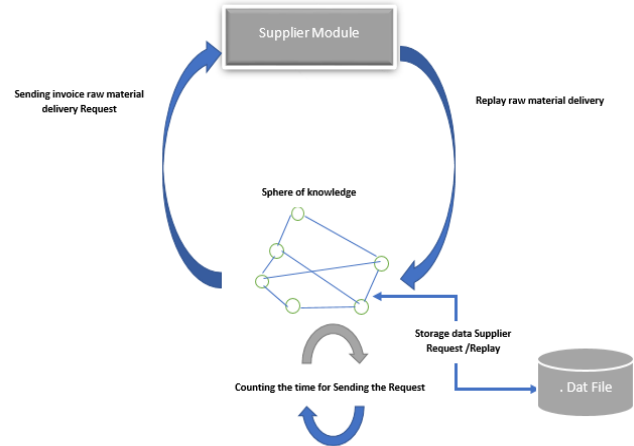


Fig. 6. Smart Supplier Process.

IV. EXPERIMENTS

From what is above, this work proves that the neural network by learning is the main engine for this intelligent system which includes the three basic modules of the company, with automation and data acquisition in real time, without errors or delay of 'sending'. This hybrid architecture dependent converged with a Java model Kohonen [9] has been planned and actualized with neuron and 26 output neurons, For the experimental simulation of the company's three modules, this work will use Any Logic software for scenario modeling and the neural Java network can work interactively with the simulation models, dynamically reading its states and taking action.

NB:

*The models have two validation times: (Validation with customer and the Internal validation) before entering in the production phase.

*The vision sensors technology of inventory administration module are simulated in the Any Logic platform as a 2/3 D bar code information transmitter.

Each module is characterized by an interactive interface linked to the Smart model company (Fig. 7, 8, 9).

In this part this work will simulate two cases of the Classic business model without neural network and the 2nd with the intelligent system on the any logic platform.

NB: In the model company with NN we just keep entering and leaving the Main model with the three modules that this intelligent system constitutes the heart which manages automatically without human intervention.

B. Delivery from Suppliers

Before the start of production, each firm needs the raw material, which implies a request from the suppliers, it requires steps to follow which can generate a lot of time to exchange calls, emails ... with a risk of errors always, the model company reduces this time, by making the entire process automated in real time from the company side to the supplier. According to statistics, the simulation present that the delivery rate to the company increases in the smart model compared to the standard model which argues that the validation period with suppliers in the classic case takes a lot of time (Fig. 14, 15).

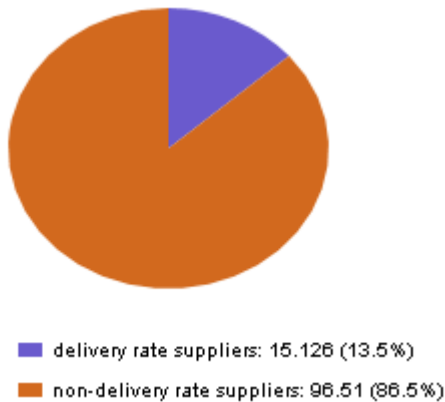


Fig. 14. Standard Delivery State of Suppliers.

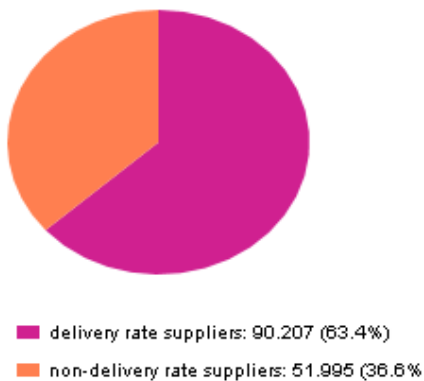


Fig. 15. Smart Delivery State of Suppliers.

C. State of Production

After validation of the products for manufacture with the customer, the demand enters a production queue, however this work run both simulation models to visualize the efficiency of the model with intelligent system. Each manufacturing process is characterized by precise planning by product, which comes first according to the degree of priority, in the classic model that most uses, we see that the products in manufacturing as well as the production line is at risk of overcapacity, it induces

a huge waiting time of products in the state ready for production, or a cancellation of the order if the time become more longer (Fig. 16, Formula 1).

In the case of introducing the intelligent system into the business model, the queue is less congested, with a rapid execution of product orders, this is due to the elimination of validation times with the customer and the time ready for internal manufacture to the company since the intelligent system made a redefinition of the refined choices with the customers and which conforms to the standards of production, which does not require an internal validation (Fig. 17, Formula 2).

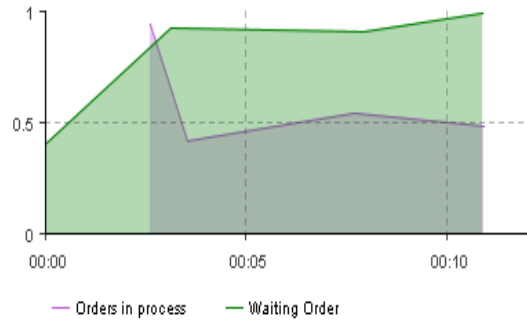


Fig. 16. Waiting Product State in Process.

Formula 1: Production start time in Standard company model

$$MS=TQ+CVT+ICVT$$

Indication's:
MS: Manufacturing start time (in progress)
TQ: Time in the queue
CVT: Customer validation time
ICVT: internal company validation time

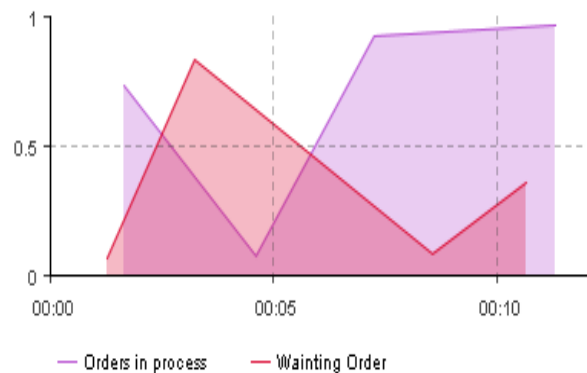


Fig. 17. Waiting Product State in Process with NN.

Nb: X axis: Duration of simulation in second

Y axis production standby and execution

Formula 2: Production start time in Smart company model

$$MS = CVTNN + TQ$$

Indication's:
MS: Manufacturing start time (in progress) CVTNN: Customer validation time NN TQ: Time in the queue

D. Time Waiting for Final Product Client

Each customer needs between his order and the reception of these requested products a definite time between telephone calls, meetings for validation of the price, quantity ...), however with the presence of the neural network the system studies the character of the customer and validates with flexibility these choices of products as previously explained. In the business model a comparative study are made of the Customer waiting time (until obtaining his order) between the classic business model and that with the neural network (Fig. 18), the statistics related to the model on the any logic platform presented as follows:

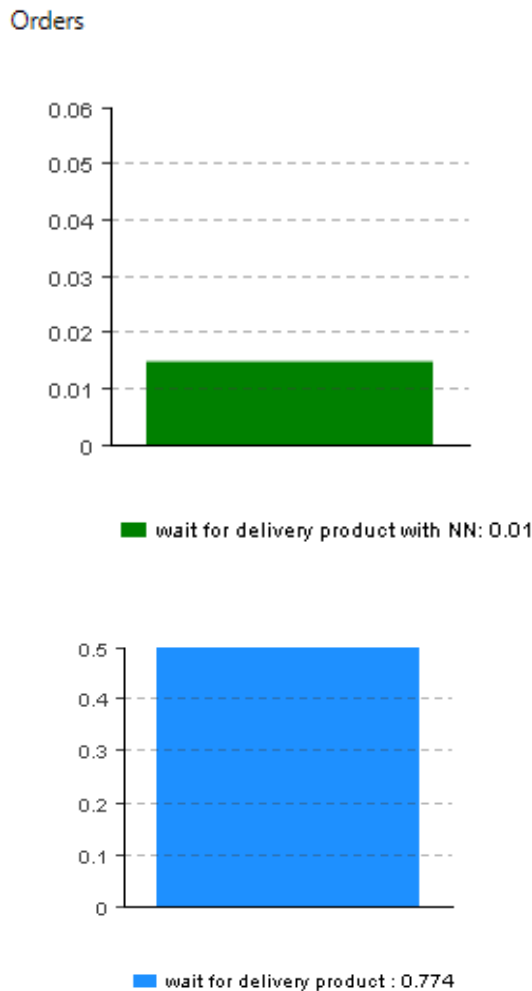


Fig. 18. Time of Waiting Client for Final Product.

Nb: The unit of measurement is the second

One of the customer's key preferences is to have their delivery on time, which is with a low probability in the case of a company with a standard system, as a solution to this problem has been designed the smart company model which reduces periods of time. Time for product validation in various phases and also reduced human resources by automating validation procedures (Formula 3, 4).

(Formula 3): Standard Company model

$$TRFD = CVT + IVT + TQ + PT + VTBD$$

Indication's:
TRFD: Time for receiving the final delivery CVT: Customer validation time IVT: Internal validation time TQ: Time in the queue PT: Production time VTBD: Validation time before delivery

(Formula 4): Intelligent Company model

$$TRFDnn = CVTnn + TQnn + PTnn$$

Indication's:
TRFDnn: Time for receiving the final delivery CVTnn: Customer validation time of NN TQnn: Time in the queue PTnn: Production time

V. CONCLUSION

Currently, organizations are gradually forced to manage their chains in a flexible and productive way by organizing and continuously administering in real time, to face the strong competition in the market, this leads them to automate their systems with new technologies, such as the intelligent system designed in this work, which serves to make the processing and transactions of the three basic business modules in real time with less human resources and error rates.

The perspective as a result of this work, will focus on the addition of a voice and facial recognition system in the business model by the eigenface and voice algorithm [10, 11], for a more secure access, and also it will integrate a fourth module concerning the personnel to record the hours, salaries and monthly bonus for each individual following a clocking system, this will allow instant up-to-date access to this personnel information, and will allow for an additional gain in productivity.

REFERENCES

- [1] Y. Yang, *C. Li, "Doppler Radar Motion Sensor With CMOS Digital DC-Tuning VGA and Inverter-Based Sigma-Delta Modulator," IEEE Trans. Instrum. Meas., vol. 63, no. 11, pp. 2666–2674, Nov. 2014.
- [2] K. Gharehbaghi*, "Artificial Neural Network for Transportation Infrastructure Systems," MATEC Web Conf., vol. 81, p. 05001, 2016.
- [3] Y. Zhang, Y.-S. Lin, I.-C. Lin, and C.-J. Chang, "Predicting logistics delivery demand with deep neural networks," in 2018 7th International Conference on Industrial Technology and Management (ICITM), Oxford, United Kingdom, 2018, pp. 294–297.
- [4] F.-M. Tsai and L. J. W. Huang, "Using artificial neural networks to predict container flows between the major ports of Asia," Int. J. Prod. Res., vol. 55, no. 17, pp. 5001–5010, Sep. 2017.

- [5] N. Karna, I. Supriana and U. Maulidevi, "Social CRM using web mining," 2014 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung, 2014, pp. 264-268, doi: 10.1109/ICITSI.2014.7048275.
- [6] N. Guo, Z. Wang, and J. Zhu, "GBVS Based 1D and 2D Barcodes Localization in Complex Scene," in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, India, 2015, pp. 352-356.
- [7] MBIDA, Mohamed. Smart Warehouse Management using Hybrid Architecture of Neural Network with Barcode Reader 1D/2D Vision Technology. *International Journal of Intelligent Systems and Applications*, 2019, vol. 11, no 11, p. 16.
- [8] O. Yorulmaz, E. Akhan, D. Tuncel, R. C. Atalay, and A. E. Cetin, "Multi-resolution super-pixels and their applications on fluorescent mesenchymal stem cells images using 1-D SIFT merging," in 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 2495-2499.
- [9] R. Thakur and L. Workman, "Customer portfolio management (CPM) for improved customer relationship management (CRM): Are your customers platinum, gold, silver, or bronze?," *J. Bus. Res.*, vol. 69, no. 10, pp. 4095-4102, Oct. 2016.
- [10] R. Rosnelly, M. S. Simanjuntak, A. Clinton Sitepu, M. Azhari, S. Kosasi and Husen, "Face Recognition Using Eigenface Algorithm on Laptop Camera," 2020 8th International Conference on Cyber and IT Service Management (CITSM), Pangkal, Indonesia, 2020, pp. 1-4, doi: 10.1109/CITSM50537.2020.9268907.
- [11] X. Zhang, Z. Tao, H. Zhao and T. Xu, "Pathological voice recognition by deep neural network," 2017 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China, 2017, pp. 464-468, doi: 10.1109/ICSAI.2017.8248337.

Handling Sudden and Recurrent Changes in Business Process Variability: Change Mining based Approach

Asmae HMAMI¹, Mounia FREDJ³
AIQualsadi Research Team, ENSIAS, University
Mohammed V of Rabat, Rabat, Morocco

Hanae SBAI²
Faculty of Sciences and Technology, University Hassan II
of Casablanca, Mohammedia, Morocco

Abstract—Changes are random and unavoidable actions in business processes, and they are frequently overlooked by managers, especially when managers need to deal with a collection of process variants. Because they must manage every single business process variant separately which is a time-consuming task. They exist many approaches to manage a collection of business process and deal with variability. Such as process mining approaches, that can discover configurable business process models, enhancing them and verify conformity automatically. However, those approaches do not cover changes and concept drift that occur over time. This paper presents a novel change mining approach that discovers changes in a collection of event logs and reports them on a change log. This change log can be analyzed to determine whether the changes are sudden or recurrent and recommend afterward some improvement to the configurable process model.

Keywords—Component; variability; process variant; configurable process model; process mining; change mining; concept drift

I. INTRODUCTION

In the last few years, there has been a growing interest in managing changes regarding the actual situation of the world related to the pandemic. Due to the coronavirus situation, many organizations must make changes to their business strategies. Education centers, schools, and universities should make online courses. Medical industries must produce a new cure. Hospitals are obliged to add more resources to support the huge demand. Other companies were required to reduce the number of employees in each space to respect the social distancing... All those changes directly affect existing models that will no longer be adapted to new conditions. In the business process context, changes are a big challenge, when changes happen during the execution stage of the business process several new features could be added and some modification can be made to the elements of the business process, so the behavior of the process is going to follow the new features rather than the existing model[1]. So those changes will reduce the validity of the first proposed models for business processes; as a result, changes must be analyzed and consider, which is an important research area on business processes. Besides, changes were managed manually with patterns and tools like ADePt (The analytical design planning technique) [2]. In order to reduce human intervention, process mining approaches are introduced to add more intelligence and automation, for model construction, validation and enhancement of processes [3]. Moreover, change management in business processes is also improved by using process mining

techniques, which is known as “change mining”. The main idea is to use event logs or change logs associated with a business process and then discover changes observed in traces [4].

However, current research on change mining focuses on discovering changes from a single event log which is associated with a single business process [4][5][6]. But companies use several copies of the same process that are similar to each other, with some differences on specific points. Those points present the variation between processes [7]. Each process variant is a business process that fits most the need that has been previously expressed by managers of this specific business. All those process variants are grouped on one model so-called configurable process model [8].

Several research works have appeared in recent years presenting and documenting configurable process models. However, to the best of our knowledge, few research works are available in the literature that addresses change mining in business process variability.

In this paper, we present a novel approach to perform change mining in a collection of event logs. The collection of event logs represents in our case a data stream that we are going to analyze. Such analyze has already been carried out by machine learning algorithms. The proposed approach in this work is inspired by one of machine learning approaches in order to extract the changed fragments from the list of events.

The remainder of this paper is organized as follows. Section 2: is a background of the most important concept related to our paper. Section 3 is a related work section. Section 4 presents the proposed change mining approach. Section 5 is an implementation and test section. Finally, Section 6 concludes the article.

II. FOUNDATIONS

To perform a change mining in a collection of event logs related to a configurable process model, it is necessary to understand the configurable process model and the variability concept. Then how the changes could impact the model.

A. Configurable Business Process Model

1) *Definition*: Companies use several copies of a business process such business process models are similar in different fragments with a slight difference in specific points. To represent and regroup all those models in one, the configurable process model concept was proposed [8].

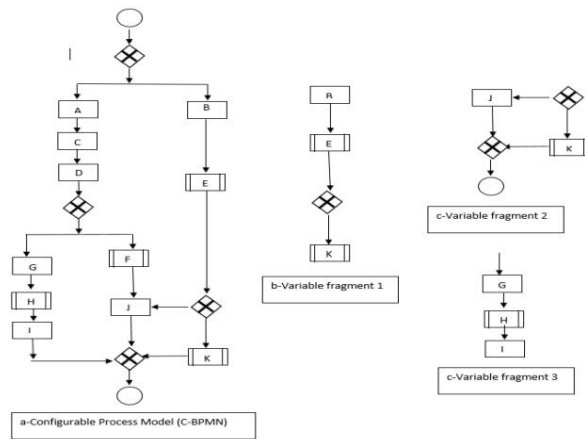


Fig. 1. Example of a Business Process Model and its Variable Fragments

Configurable process models aim to provide generic models integrating possible process variations into one model. Afterward such a model can be configured to a specific solution. This means a configurable model should guide the user to a solution that fits with the user's requirements. [8]

The configurable business process model can be represented by some specific modeling language such as C-EPC [7], C-BPMN [9] The modeling language must provide the needed options that can help to create and drive desired process variants from a configurable process model.

2) *Key elements*: The important concept of a configurable process model is "variability", which is related to two elements: Variation point and Variants.

- The variation points are locations likely to be different in each business process variant [8].
- The variants are the possible values that variation points can have in each process variant [8].

A variable fragment is a subset of a business process model that captures variability [10]. It contains at least one variation point.

As an example, part (a) of the Fig. 1 presents a configurable process model and parts b) c) d) in the same figure are some variable fragments of this specific model.

When the configuration is made based on the model, different possibilities are available, and choices can be made from the list of variants of each variation point, to create the desired business process variant. According to the "hide and block" technique [9], from the list of variants each specific configuration choices to ON (use the variation) OFF (hide the variation) OPT (The choice depend on some condition) the variation, so we have to use none, one or many variants for each variation point [9].

3) *Creation of a configurable process model*: To create a configurable business process model there are two different methods based on whether it will be created from scratch or by using a process mining technique.

a) *Creation from scratch*: Configurable process models can be constructed in different ways. They can be designed

from scratch, but if a collection of existing process models already exists, a configurable process model can be derived by merging the different variants. [11]

Different approaches have been proposed in order to merge existing process models into a configurable process model. [12][13][14][15] the input of almost all those approaches is a collection of business process models of the same family and the output is a configurable process model.

b) *Creation by using process mining*: Another way of obtaining a configurable process model is not by merging process models but by applying process mining techniques on a collection of event logs.

The aim of process mining is to use the recorded data about the previous execution stored in a file called event logs [3] in order to discover new models, enhance business processes or verify conformity [3].

There are four approaches for discovering configurable process models. [16]

- Merging individually discovered process models: the configurable process model is discovered based on merging the discovered process variants which is made individually.
- Merging similar discovered process models from a common model: in this approach, the configurable business process is discovered based on merging discovered process variants which are made from the discovered common model.
- Discovering a single process model then discover configurations: in this approach, a common model is discovered from the collection of event logs and secondly, the configuration is mined from the common model in order to create a configurable process model.
- Discovering process model and configurations at the same time: in this approach, the configurable process model is created directly from the collection of event logs [16].

However, process mining techniques do not take into account changes that may occur during the life cycle of the process. Because business processes are not in steady state, so the configurable process model discovered from event logs, can no longer be adapted to the real situation.

In the next section, we define changes in variability. We start by defining changes and when they will happen, then how do they show up on the event log.

B. Changes in Business Process Variability

Changes in variability can be classified into two categories predictable changes and unpredictable ones.

1) *Predictable Change*; (Reengineering, redesign, improvement)

Each configurable process model once created may be subject to many changes during its life cycle, due to different circumstances. Those changes can be related to a reengineering or a redesign or an improvement ... [17].

The change in the business process management field has various definitions. We selected below the most relevant ones:

- Change is the fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service, and speed [17].
- business process change management is a strategy-driven organizational initiative to improve and (re)design business processes to achieve competitive advantage in performance (e.g., quality, responsiveness, cost, flexibility, satisfaction, shareholder value, and other critical process measures) through changes in the relationships between management, information, technology, organizational structure, and people[18].

As mentioned above, those changes are performed and happen during each phase of the life cycle of the business process model, in each phase a specific modification can be made, which is depicted into four phases:

- Phase (a) : process design, which is the first step and in this step, changes can be made as a redesign or reengineer,
- Phase (b): process configuration, in this phase process variant, is created from a configurable process model by choosing for each variation point one or many variants, in this phase, we can add implicitly a variant or a variation point depending on the new requirements,
- Phase (c): process enactment, the process variant is made into production and test in order to verify the compliance with the need, if minor adjustments are required, they will be made. So changes will be made on the model,
- Phase (d): a process diagnosis phase which leads to process adaptations, and in this phase, we can recommend a new model, to design new process models [19].

However, on the one hand, this type of change is not all reported in guidelines, in order to solve a problem quickly, many managers can perform changes without documenting the performed actions, which can lead to some confusion when working on the same business process model. On the other hand, predictable changes are not the only cause of changes. Concept drift can also change the behavior and the structure of a configurable process model.

2) Unpredictable change: Concept Drift

The configurable process model can during its life cycle, meet some unpredictable changes that are not made by managers but occur due to some actions made by other users or systems running. Those changes are named concept drift, this type of change will affect the initial concept (which is subject to change). There are four types.

- Sudden change: an unanticipated event that occurs or takes place unexpectedly,

- Recurring change: seasonal changes, that appear many times over time,
- Gradual change: this Change starts with a limited context and increase slowly to be finally applied to the entire stream,
- Incremental change: small different mutations happen to the concept many times until it becomes a new completely different concept [20].

In this paper, we are concerned with the first two types of changes.

C. Definition of Variability Change in a Collection of Event Logs

1) *Definition*: The presented changes are almost all related to the execution of the business process and are related to the behavior of the business process and how it is executed. And as the execution is recorded on event logs, we will search changes from event logs which is the dynamic aspect of the business process.

In variability context, we are using not only one event log but a collection of event logs.

So, a changed event in a collection of event logs is an event that occur multiple times in the event log, and this event is different from all the possible events from all the possible process variants of the same family.

From this definition, an event (in the context of process variability) can be concerned as a changed one if and only if:

- The event is repeated in many traces of the same process variant (if not those events are concerned as errors).
- The event is not expected. Not only in the process variant where the change happens but also in the other process variants of the same family.

To illustrate variability changes and the importance of detecting those changes let us take look at the example in the next sub-section.

2) *Example of change in a collection of event logs*: From a configurable business process model, many process variants are driven and during their execution, events go through activities of the business process to compose a trace. From each complete execution, those traces are recorded on event logs. This event log is the input of process mining techniques. However, the recorded traces do not fit all the normal behavior that is described in the process model. If these unexpected behaviors are not detected and labeled, they will lead to errors when performing process mining.

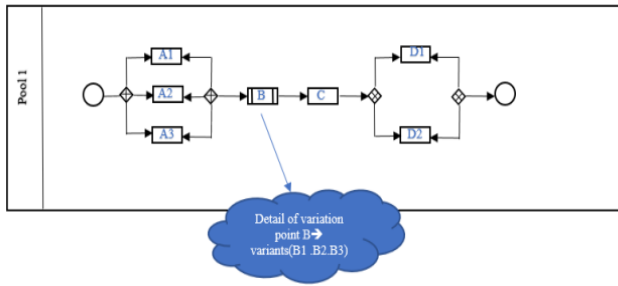


Fig. 2. Example of a Configurable Process Model.

Fig. 2 is an example of a configurable process model. This model can generate three process variants. The recorded traces can have the following variability changes.

- <A1B3C1D2D1> 1- Normal trace
- <A2ZC1D2D1> 2-Trace with change on the variation.
- <A3B3ND2D1> 3-Change on the variable fragment.
- <A1B3A1B3C1D2D1> 4- Change in the order of execution.

- Normal trace: is our reference and we consider it is normal because it follows the structure and the behavior of the predefined configurable process model.
- Trace with change in the variation: in this trace, the variant is different from the list of variants of this specific variation point.
- Change on the variable fragment (activity): the directed connected activities to the variants have changed.
- Change on the variable fragment (execution): the sequence flow has changed due to the change in the order of the execution of activities.

To highlight the importance of detecting changes before applying a process mining technique on the event log, we have applied a process mining algorithm on event logs that contained the list of changes presented in the example. The obtained result is presented in Table 1, which presents the discovered business process model obtained after applying Alpha algorithm [21] exiting in prom as plugin [22].

As we can observe in Tab.1 the discovered model contains activities that have been added or removed due to changes, and models are no more the same as it is in the predefined one. We can easily recognize that without hiding those changes and identifying them, the process mining algorithm will lead to confusion.

TABLE I. EXAMPLE OF PROCESS MINING IN LOGS WITH CHANGES

Applied changes	Trace Example	Discovered model
No changes	<A1B3C1D2D1> <A2B3C1D1D2> <A3B3C1D2D1>	
Change in variants	<A1B3C1D2D1> <A2NBC1D1D2> <A3NBC1D2D1> <A1NBC1D2D1> <A2B3C1D2D1>	
Change in activities on the variable fragment	<A1B3C1D2D1> <A2B3NND1D2> <A3B3NND2D1> <A1B3C1D2D1> <A2B3C1D1D2> <A3B3NND2D1>	
Change in sequence of the variable fragment	<A1B3A1B3C1D2D1> <A2B3A2B3C1D1D2> <A3B3A3B3C1D2D1>	

III. RELATED WORK

Managing changes has been widely discussed in the field of business process management, many works have proposed approaches to deal manually with changes such as AdePt [2].

In recent years change mining, which is an automatic approach to discover changes has emerged and those approaches were widely used to detect changes in business processes. In Previous work, a comparative study has been conducted [23]. This study shows that almost all selected papers in this comparative study had dealt with changes in a single business process [4][5][6]. And only a few papers have proposed approaches to deal with changes in a collection of business processes [24][25]. Those approaches are limited to propose some rules [24] related specially to the configuration and how process variants could be created from the configurable process model based on the observed behaviors [25]. However, they do not detect changes over time which are known as concept drift.

Recently work on concept drift in the business process have taken some importance and novel approaches have proposed [26][27][28] to detect concept drifts that occur during the execution phase of the business process. However, none of them have dealt with the variability concept.

So, our goal in this paper is to overcome the limitation of change mining approaches related to variability by proposing an approach that can detect changes in variability. The proposed approach is based on mining change in data recorded on event logs of a collection of process variants. Afterward, the detected changes will be reported in change log of variability that can be used in future work to recommend improvement to the configurable process model.

IV. PROPOSED CHANGE MINING APPROACH

As it is presented in our previous work, the proposed approach is based on many steps that are presented in the figure below (Fig. 3).

A. Steps of the Proposed Change Mining Approach

- Preparing event log of variable fragments: In this step, we prepare the event log of variable fragments using the merging and the filtering approach presented in our previous work [29]. This step's starting point is a variability specification file, a collection of event logs, and the output is the event log of variable fragments.
- Applying Change mining algorithm: this step is based on mining change from the event log of variable fragments and creating the change log. The detail of this algorithm is presented in this paper.
- Analyzing and recommending operations: in this phase, some metrics are used to identify the most significant changes to recommend as a future evolution of the configurable business process model.

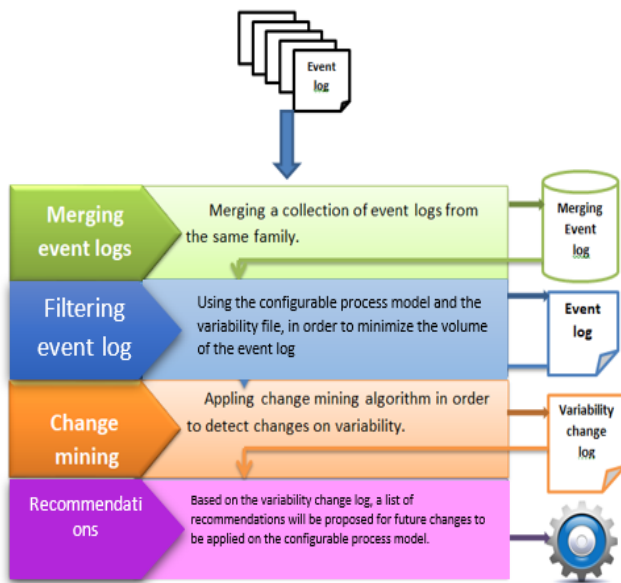


Fig. 3. Change Mining Framework.

B. Change Mining Algorithm

The proposed change mining algorithm is performed on the event log of variable fragments which is the output of merging and filtering algorithms proposed in the previous work [29]. So in this work, the input is an event log of variable fragments that contains events of variable fragments. The proposed algorithm is based on six steps: Classification, Initialization, Projection, Evaluation, Aggregation, Storage, Fig. 4.

Those steps are inspired by an existing algorithm to detect concept drift in a data stream called STAGGER. This algorithm is one of the first ones used in machine learning to overcome the problem of concept drift [30]. It is based on attributing three discrete attributes with three possible values each, for example

- $size \in \{small, medium, large\}$
- $color \in \{green, blue, red\}$
- $shape \in \{triangle, circle, rectangle\}$ [31].

In order to project the STAGGER algorithm description into our case, let first assume that an event log is an ensemble A that contains vectors T, each vector is a trace of the event log. Each trace is a collection of events grouping a certain number of activities. So $A = \sum (T)$.

The event log of variable fragments is ensemble E contains many events log of each process variant of A without the common elements. If B is an ensemble of the common element the ensemble E is $E = \sum (A \cap B)$.

And finally, the variable fragment is a vector F of three elements.

So, the three discrete attributes are process variant, fragments, and fragment's elements. In the STAGGER algorithm for each attribute, they are three values however in our case the number of possible values is a number greater than one.

- The number of possible values for process variants depends on the number of event logs in the collection.
- The number of values for fragments depends on the number of variation points in the configurable process model.
- Values for fragment's elements will be a vector of three components that present the previous, the variation and the next elements. The possible parameter for each value of the attribute fragment's elements are previous (start point, activity, or list of activity) a variation (activity) next (activity, list of activity or end point).

Thus, our discrete attribute will have the following values

- process variant $\in \{\text{all possible process variants}\}$,
- fragments $\in \{\text{all possible fragments}\}$,
- fragment's element $\in \{\text{previous, variation, next}\}$,

And for example, a fragment will have the following syntax $\langle PV1, F3, [A, M1, B] \rangle$.

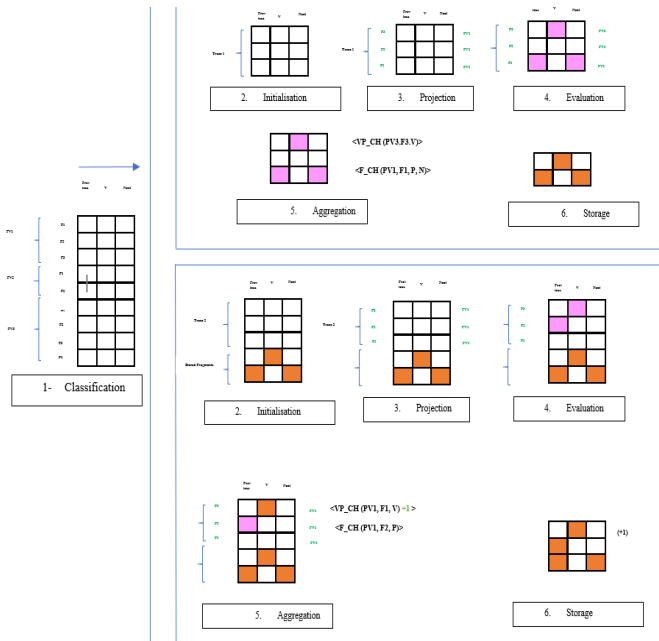


Fig. 4. Behavior of the Change Mining Algorithm.

Those discreet attributes with their values are the keys elements of each step of the proposed algorithm, in each step; we will have some tasks to complete as they are described on Fig. 4, in order to get after many iterations through the event log of variable fragments, a change log.

Tasks in each step are as follows.

- **Classification:** labeling a list of fragments that will be used to select valid and invalid fragments when we loop on the event log of variable fragments. In this step, each attribute's values will be assigned, and a list of valid fragments will be created. Fig. 5 presents the algorithm for this phase.
- **Initialization:** initialization of the fragment pool with fragments of the trace (i). These fragments will be formatted in the form of the chosen attribute (Fig. 6).
- **Projection:** projection of the selected trace's fragments on the list of valid fragments created on the classification step. This projection finds the most likely fragments to the selected fragment (Fig. 6).
- **Evaluation:** in this step, if the projection detects changes in one of trace's fragments, we will identify the specific element concerned by the change. If change is in previous or next elements, we have a position change, which is fragment change. If the change is on the variation, it is a change in variation points (Fig. 6).
- **Aggregation:** names the change by its appropriate target (Change on fragments, change on variants, or change on variation point) and gives a count number to the target name to put it on a specific change type group (Fig. 6).
- **Storage:** deletes fragments that did not meet a change from the pool and store the changed one on an XML file (Fig. 6).

Step 1: classification

```

1-while (variability_file) do
2-For (i=1; i<=nombre_previous; i++)
3-    For (j=1; j<= nombre_variants; j++)
4-        For (k=1 ; k<= nombre_next; k++)
5-            Var_new= Previous[i]. variants[j].next[k]
6-            Store (Var_new)
7-        End for
8-    End for
9-End for
10-End while
    
```

Fig. 5. Classification Algorithm.

Steps 2, 3, 4, 5, 6:

```

1. While (target_event_log_of_variable_fragments) do
2. //Step 2: initialization
3. Array M_trace //while M_trace is a matrix variable fragments
4. Array F //while F is a vector for each fragment.
5. For (each fragment in trace)
6.     F[process_variant]=trace_id;
7.     F[fragments]=fragment_id;
8.     F[Fragment_elements] = [fragment[0], fragment[1], fragment[2]] ;
9.     M_trace[]=F ;
10. End for
11. //Step 3: Projection
12. For(F[i] of the new trace of M_trace)
13.     If(F[i] is not stored_fragments)
14.         F[change][i]='change';
15. //Step 4: Evaluation
16.     F_near=Equidistance(stored_fragments,F[i]);
17.     Variant=Compare_diff_variant(F_near,F[i]);
18.     Previous=Compare_diff_previous(F_near,F[i]);
19.     Next=Compare_diff_next(F_near,F[i]);
20. //Step 5: Aggregation
21.     F[targed][i]=[previous, variant,next];
22.     End if
23. End for
24. //Step 6: Storage
25. For(F[i] of the new trace of the M_trace)
26.     If(F[change][i]=="null")
27.         Discard(F[i])
28.     Else
29.         Store(F[targed][i])
30.     End for
31. End while
    
```

Fig. 6. Initialization, Projection, Evaluation, Aggregation, Storage Algorithms.

The first algorithm in Fig. 5 is for the initialization step. The second sub-code in the Fig. 6 is for steps that we will loop on through the event log of variable fragments

At the end of the proposed algorithm, we will have a change log as an output. This change log is formatted as XML format and contains a chronologically sorted list of detected changes. Fig. 7 is an example of the obtained change log.

```

<?xml version="1.0" encoding="UTF-8" ?>
<change>
  <id>23</id>
  <date>2019-01-01T16:03:01.012345</date>
  <tarce_id>56</tarce_id>
  <id_pv>2</heading>
  <id_variation_point>5</id_variation_point>
  <type>next_fragment_change</type>
  <change_true>Activity_B</change_true>
  <change_old>Activity_W</change_old>
</change>
<change>
  <id>23</id>
  <date>2019-01-01T16:04:08.0325468</date>
  <tarce_id>20</tarce_id>
  <id_pv>3</heading>
  <id_variation_point>4</id_variation_point>
  <type>variant_change</type>
  <change_true>Activity_M</change_true>
  <change_old>Activity_Z</change_old>
</change>
    
```

Fig. 7. Sub-Part of a Change Log obtained with the Proposed Algorithm.

The generated change logs should be detailed enough and accurate enough, to provide the information required for performing a future analyzes.

It must answer the following question

- 1) When did the change happens?
- 2) Which trace is concerned by this change?
- 3) Which business process variant is concerned by this change?
- 4) Which variability is concerned?
- 5) It is a fragment change or variants change?
- 6) Which is the name of the concerned element by this change?
- 7) What is the new element in this change?

In order to answer this list of questions each detected change is stored in the xml file with the following attribute:

- 1) Date of change.
- 2) Id of trace.
- 3) Id of process variant
- 4) Variation point
- 5) Type of change (Fragments or variants).
- 6) Name of the changed element
- 7) Name of the new value due to the change.

V. IMPLEMENTATION, PROTOTYPE, AND TESTS

A. Implementation and Prototype

The proposed approach is implemented as a new function on the toolset, the “random configurable process model generator” [32].

This toolset is a set of functions that provide the ability to generate random business process models with their process variants and simulate their execution and get event logs. Implementing the change mining algorithm in this tool will facilitate to test the implemented algorithm because all the required inputs are available in the same tool.

The algorithm takes as input only the event log of variable fragments. Because we implemented the algorithm in the same environment where previous algorithms have already been implemented, which are filtering and merging ones.

However, it is possible to run all three functions as one, if we have all the required input which are the variability specification file and a collection of event logs from the interface shows in Fig. 8.

Our toolset gives to the user the ability to choose between having an event log with or without changes. The user can also choose the type of change to apply from the interface presented in Fig. 9. This will help us to test our change mining algorithm on event logs with different types of changes.

To test the proposed algorithm, we use a running example of collections of event logs based on three business process models. The three models are generated randomly using the toolset.

- Model 1: contains two variation points. The first variation point has 3 variants and the second has 4 variants Fig. 10.
- Model 2: contains four variation points. The first variation point has 3 variants and the second has four variants the third has three variants and the fourth has four variants Fig. 11.
- Model 3: contains six variation points. The first variation point has 3 variants and the second has four variants, the third has three variants, the fourth has four variants the fifth has three variants and finally the sixth has four variants Fig. 12.

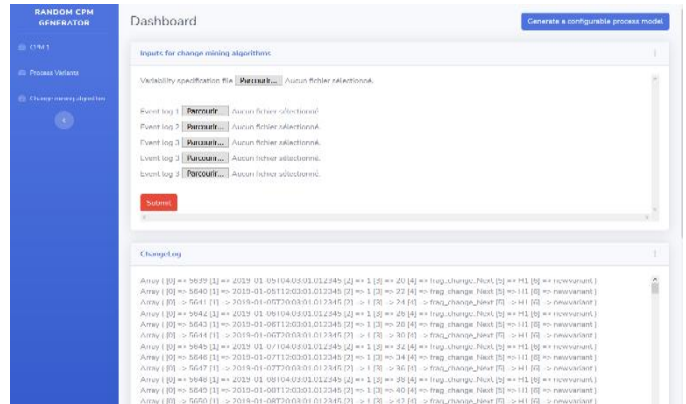


Fig. 8. Interface for Collecting Inputs and Performing a Change Mining.

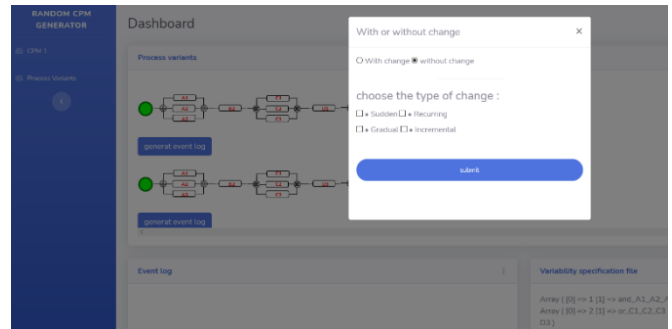


Fig. 9. Interface for Choosing the Type of Change to Add Randomly to the Event Log.

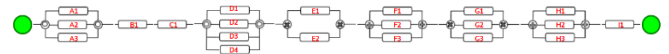


Fig. 10. Configurable Process Model 1.



Fig. 11. Configurable Process Model 2.

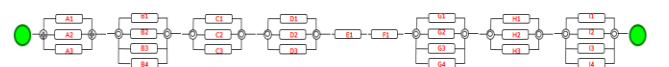


Fig. 12. Configurable Process Model 3.

From the three generated business process model, we will generate a collection of process variants each collection. The toolset will simulate the execution of those process variants and generate their event logs. In addition, we will apply randomly on the obtained event logs a different type of drift to get an event logs with changes.

B. Tests and Results

The three collections of event logs are shown below.

- Collection 1: contains three event logs of three process variants of a configurable business process model 1 (Fig. 10). Each event log contains 100 traces.
- Collection 2: contains six event logs of six process variants of a configurable business process model 2 (Fig. 11). Each event log contains 100 traces.
- Collection 3: contains twelve event logs of twelve process variants of a configurable business process model 3 (Fig. 12). Each event log contains 100 traces.

When generating event logs, we will apply random changes to each event log.

For each collection, we perform a change mining in order to detect the applied changed on the event log and generate the change log. First, we apply the merging and filtering algorithms to get the event log of variable fragments. Fig. 13(a) is a subpart of the event log of variable fragment and we highlight changed fragment with a green shape. Second, we apply the change mining algorithm. Finally, we export the results of the mining as an XML file, which is the change log. An example of the obtained change log is presented in Fig. 13(b).

We made the same actions on the three collections and in each collection, we were able to generate change log and detect all most all applied changes.

```
5643,2019-01-06T15:03:01.012345,C1,3,387
5643,2019-01-06T16:03:01.012345,newvariant,3,387
5643,2019-01-06T17:03:01.012345,E1,3,387
5644,2019-01-06T22:03:01.012345,A3,3,388
5644,2019-01-06T23:03:01.012345,B1,3,388
5644,2019-01-07T00:03:01.012345,C1,3,389
5644,2019-01-07T01:03:01.012345,newvariant,3,389
5644,2019-01-07T02:03:01.012345,E2,3,389
5645,2019-01-07T07:03:01.012345,A3,3,390
5645,2019-01-07T08:03:01.012345,B1,3,390
5645,2019-01-07T09:03:01.012345,C1,3,391
5645,2019-01-07T10:03:01.012345,newvariant,3,391
5645,2019-01-07T11:03:01.012345,E1,3,391
5646,2019-01-07T16:03:01.012345,A3,3,392
5646,2019-01-07T17:03:01.012345,B1,3,392
5646,2019-01-07T18:03:01.012345,C1,3,393
5646,2019-01-07T19:03:01.012345,newvariant,3,393
5646,2019-01-07T20:03:01.012345,E1,3,393
5647,2019-01-08T01:03:01.012345,A3,3,394
5647,2019-01-08T02:03:01.012345,B1,3,394
5647,2019-01-08T03:03:01.012345,C1,3,395
5647,2019-01-08T04:03:01.012345,newvariant,3,395
5647,2019-01-08T05:03:01.012345,E1,3,395
5648,2019-01-08T10:03:01.012345,A3,3,396
5648,2019-01-08T11:03:01.012345,B1,3,396
5648,2019-01-08T12:03:01.012345,C1,3,397
5648,2019-01-08T13:03:01.012345,newvariant,3,397
5648,2019-01-08T14:03:01.012345,E1,3,397
```

```
<?xml version="1.0" encoding="UTF-8"?>
<change>
  <id>1</id>
  <date>2019-01-05T10:03:01.012345</date>
  <trace>5639</trace>
  <pv>2</pv>
  <fragment_id>179</fragment_id>
  <change_type>change_variations</change_type>
  <old_element>D1</old_element>
  <new_element>newvariant</new_element>
</change>
<change>
  <id>2</id>
  <date>2019-01-05T19:03:01.012345</date>
  <trace>5640</trace>
  <pv>2</pv>
  <fragment_id>181</fragment_id>
  <change_type>change_variations</change_type>
  <old_element>D1</old_element>
  <new_element>newvariant</new_element>
</change>
<change>
  <id>3</id>
  <date>2019-01-06T04:03:01.012345</date>
  <trace>5641</trace>
  <pv>2</pv>
  <fragment_id>183</fragment_id>
  <change_type>change_variations</change_type>
  <old_element>D1</old_element>
  <new_element>newvariant</new_element>
</change>
</change>
```

(a) The input: event log of variable fragments. (b) The output: change log.

Fig. 13. Input and Output of the Change Mining Algorithm.

VI. CONCLUSIONS

This paper presents a novel approach to perform a change mining in a collection of event logs based on a modified STAGGER algorithm.

Our approach is based on detecting sudden and recurrent change by using steps of STAGGER algorithm and storing detected changes in an XML file so-called change log.

The proposed approach is implemented on the toolset “random configurable process model generator”, and it shows its ability to detected drift on synthetic event logs.

In this work, we are concerned only by sudden and recurrent changes. However, more improvement can be applied to detect the other types of changes.

We also aim to test our proposed approach on a real collection of event logs and add more change mining perspectives especially data and resources.

As future work, we aim to create from the generated change log, a recommendation system that proposes a new configurable process model based on the detected changes. Also, as perspective, we intend to make our approach suitable with the situation where the configurable process model is not discovered.

REFERENCES

- [1] Song, W., & Jacobsen, H. A. (2016). Static and dynamic process change. *IEEE Transactions on Services Computing*, 11(1), 215-231.
- [2] Austin, S., Baldwin, A., Li, B., & Waskett, P. (2000). Analytical design planning technique (ADePT): a dependency structure matrix tool to schedule the building design process. *Construction Management & Economics*, 18(2), 173-182.
- [3] Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... & Wynn, M. (2011, August). Process mining manifesto. In *International Conference on Business Process Management* (pp. 169-194). Springer, Berlin, Heidelberg.
- [4] Hompes, B., Buijs, J. C., van der Aalst, W. M., Dixit, P. M., & Buurman, H. (2015). Detecting Change in Processes Using Comparative Trace Clustering. *SIMPDA*, 2015, 95-108.
- [5] Kaes, G., & Rinderle-Ma, S. (2017, June). On the similarity of process change operations. In *International Conference on Advanced Information Systems Engineering* (pp. 348-363). Springer, Cham.
- [6] van der Aalst, W. M. (2015, August). Change Point Detection and Dealing with Gradual and Multi-order Dynamics in Process Mining. In *Perspectives in Business Informatics Research: 14th International Conference, BIR 2015, Tartu, Estonia, August 26-28, 2015, Proceedings* (Vol. 229, p. 161). Springer.
- [7] van der Aalst, W. M., Dreiling, A., Gottschalk, F., Rosemann, M., & Jansen-Vullers, M. H. (2005, September). Configurable process models as a basis for reference modeling. In *International Conference on Business Process Management* (pp. 512-518). Springer, Berlin, Heidelberg.
- [8] Gottschalk, F., Van der Aalst, W. M., & Jansen-Vullers, M. H. (2007). Configurable process models—a foundational approach. In *Reference Modeling* (pp. 59-77). Physica-Verlag HD.
- [9] Zhang, H., Han, W., & Ouyang, C. (2014, September). Extending BPMN for Configurable Process Modeling. In *ISPE CE* (pp. 317-330).
- [10] Mancioffi, M., Danylevych, O., Karastoyanova, D., & Leymann, F. (2011, August). Towards classification criteria for process fragmentation techniques. In *International Conference on Business Process Management* (pp. 1-12). Springer, Berlin, Heidelberg.
- [11] Buijs, J. C., van Dongen, B. F., & van der Aalst, W. M. (2013). Mining configurable process models from collections of event logs. In *Business process management* (pp. 33-48). Springer, Berlin, Heidelberg.
- [12] Gottschalk, F., van der Aalst, W.M.P., Jansen-Vullers, M.H.: Merging Event-driven Process Chains. In: Meersman, R., Tari, Z. (eds.) OTM 2008, Part I. LNCS, vol. 5331, pp. 418–426. Springer, Heidelberg (2008)
- [13] La Rosa, M., Dumas, M., Uba, R., Dijkman, R.: Business Process Model Merging: An Ap-proach to Business Process Consolidation. *ACM Transactions on Software Engineering and Methodology* 22(2) (2012)

- [14] Schunselaar, D.M.M., Verbeek, E., van der Aalst, W.M.P., Raijers, H.A.: Creating Sound and Reversible Configurable Process Models Using CoSeNets. In: Abramowicz, W., Kriksciuniene, D., Sakalauskas, V. (eds.) BIS 2012. LNBP, vol. 117, pp. 24–35. Springer, Heidelberg(2012)
- [15] Gottschalk, F., van der Aalst, W.M.P., Jansen-Vullers, M.H.: Mining Reference Process Models and their Configurations. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2008.LNCS, vol. 5333, pp. 263–272. Springer, Heidelberg (2008);
- [16] Buijs, J.C.A.M., La Rosa, M., Reijers, H.A., Dongen, B.F., van der Aalst, W.M.P.: Improving Business Process Models using Observed Behavior. In: Proceedings of the Second International Symposium on Data-Driven Process Discovery and Analysis. LNBP, Springer (toappear, 2013).
- [17] Kettinger, W. J., & Grover, V. (1995). Toward a theory of business process change management. *Journal of Management Information Systems*, 12(1), 9-30.].
- [18] Kettinger, W.J.; Guha, S.; and Teng, J.T. The process reengineering life cycle methodology: a case study. In V. Grover and W.J. Kettinger (eds.), *Business Process Change: Reengineering Concepts, Methods and Technologies*. Hanisburg, PA: Idea Group Publishing, 1995
- [19] Gottschalk, F. (2009). Configurable process models.
- [20] Kadam, S. (2019). A Survey on Classification of Concept Drift with Stream Data.
- [21] W. van der Aalst, T. Weijters, and L. Maruster, “Workflow mining: Discovering process models from event logs,”*IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.
- [22] Van der Aalst, W. M., van Dongen, B. F., Günther, C. W., Rozinat, A., Verbeek, E., & Weijters, T. (2009). ProM: The process mining toolkit. *BPM (Demos)*, 489(31), 2.
- [23] Hmami, A., Sbai, H., & Fredj, M. (2019, April). Change Mining in Business Process Variability: A Comparative Study. In 2019 4th World Conference on Complex Systems (WCCS) (pp. 1-5). IEEE.
- [24] Assy, N., & Gaaloul, W. (2014, November). Configuration rule mining for variability analysis in configurable process models. In *International Conference on Service-Oriented Computing* (pp. 1-15). Springer, Berlin, Heidelberg.
- [25] Arriagada-Benítez, M., Sepúlveda, M., Munoz-Gama, J., & Buijs, J. C. (2017). Strategies to automatically derive a process model from a configurable process model based on event data. *Applied sciences*, 7(10), 1023.
- [26] Baier, L., Reimold, J., & Kühl, N. (2020, June). Handling Concept Drift for Predictions in Business Process Mining. In 2020 IEEE 22nd Conference on Business Informatics (CBI) (Vol. 1, pp. 76-83). IEEE.
- [27] Bose, R. J. C., van der Aalst, W. M., Žliobaitė, I., & Pechenizkiy, M. (2011, June). Handling concept drift in process mining. In *International Conference on Advanced Information Systems Engineering* (pp. 391-405). Springer, Berlin, Heidelberg.
- [28] Yeshchenko, A., Di Ciccio, C., Mendling, J., & Polyvyanyy, A. (2021). Visual Drift Detection for Sequence Data Analysis of Business Processes. *IEEE Transactions on Visualization and Computer Graphics*.
- [29] Hmami, A., Sbai, H., & Fredj, M. (2020, March). A new Framework to improve Change Mining in Configurable Process. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security* (pp. 1-6).
- [30] Ceravolo, P., Tavares, G. M., Junior, S. B., & Damiani, E. (2020). Evaluation goals for online process mining: a concept drift perspective. *IEEE Transactions on Services Computing*.
- [31] Schlimmer, J. C., & Granger, R. H. (1986, August). Beyond Incremental Processing: Tracking Concept Drift. In *AAAI* (pp. 502-507).
- [32] Hmami, A., Sbai, H., & Fredj, M. (2020, March). Enhancing change mining from a collection of event logs: Merging and Filtering approaches; *Journal of Physics: Conference Series*, Volume 1743, 2021 - IOPscience.

The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem

Mustafa Abdul Salam¹

Artificial Intelligence Dept

Faculty of Computers and Artificial Intelligence, Benha
University, Benha, Egypt

Ahmad Taher Azar²

Faculty of Computers and Artificial Intelligence, Benha,
University, Egypt. and College of Computer and Information
Sciences, Prince Sultan University, Riyadh, Kingdom of
Saudi Arabia

Mustafa Samy Elgendy³

Scientific Computing Dept.

Faculty of Computers and Artificial Intelligence, Benha
University, Benha, Egypt

Khaled Mohamed Fouad⁴

Information Systems Dept

Faculty of Computers and Artificial Intelligence, Benha
University, Benha, Egypt

Abstract—In most conditions, it is a problematic mission for a machine-learning model with a data record, which has various attributes, to be trained. There is always a proportional relationship between the increase of model features and the arrival to the overfitting of the susceptible model. That observation occurred since not all the characteristics are always important. For example, some features could only cause the data to be noisier. Dimensionality reduction techniques are used to overcome this matter. This paper presents a detailed comparative study of nine dimensionality reduction methods. These methods are missing-values ratio, low variance filter, high-correlation filter, random forest, principal component analysis, linear discriminant analysis, backward feature elimination, forward feature construction, and rough set theory. The effects of used methods on both training and testing performance were compared with two different datasets and applied to three different models. These models are, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest classifier (RFC). The results proved that the RFC model was able to achieve the dimensionality reduction via limiting the overfitting crisis. The introduced RFC model showed a general progress in both accuracy and efficiency against compared approaches. The results revealed that dimensionality reduction could minimize the overfitting process while holding the performance so near to or better than the original one.

Keywords—Dimensionality reduction; feature subset selection; rough set; overfitting; underfitting; machine learning

I. INTRODUCTION

Overfitting could be defined as the curse of a machine learning classifier and would probably be considered as the most common problem for beginners. It was a challenging problem with enthralling solutions that lied in dealing with the procedure's arrangements. Overfitting was an essential trouble which appeared illogically from outside; it occurred when the model proved its data accurately. [1][2].

The only service of the leaning-difference crisis was for observing when the model stepped into underfitting or overfitting. This Bias variance trouble is basic for a guarded machine learning. It's a method to identify the outcomes of the algorithm via dividing the evaluation error down. There are three kinds of error to be expected:

1) *Bias error*: The bias error was calculated by indicating the difference between the model's predicted evaluation and the real value that the model had been testing to reach.

2) *Variance error*: According to a certain data point of view, the variance error came from the turbulence of a sample predictions.

3) *The irreducible error*: It was likely to find out overfitting and underfitting Since training error and non-test error were gained by the low overfitting results. On the contrary, underfitting led to great training and a collection of test errors, as shown in figure 2 below.[2][3].

However, overfitting occurred in case that the model matched with the data very well, as illustrated in figure 1 (a). Underfitting took place whenever the model or algorithm was not applied to the data typically as cleared in figure 1 (b).

By investigation, many ways were reached to skip the overfitting

1) *Regularization*: In machine learning, regularizing the criterion was regarded as one technique in order to reduce, regularize, or narrow the coefficient value into zero. Such a method hindered exploring a more elaborated or even an all-purpose model. It suppressed the occurrence of overfitting, as revealed in Figure 3.

- Greenline reflected the coefficient before regularization
- Blue Line conveyed the coefficient after regularization

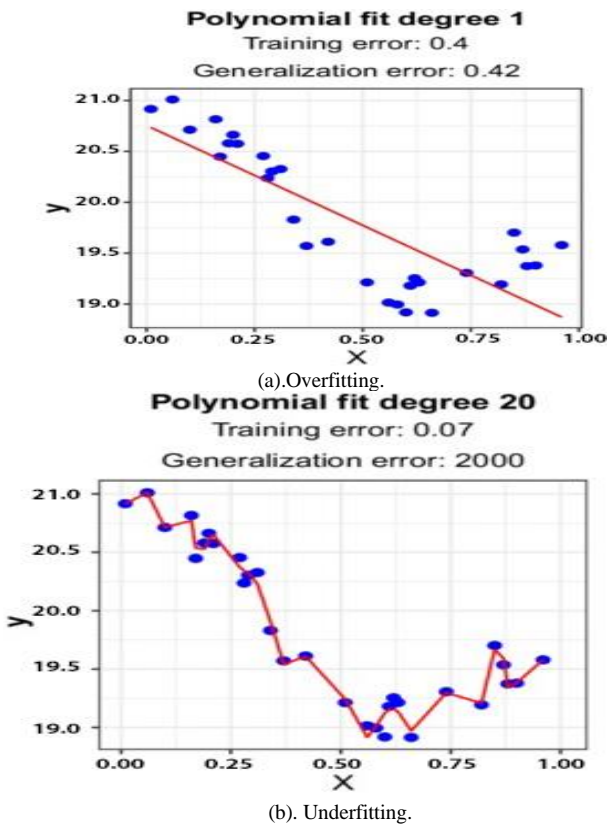


Fig. 1. Overfitting and Underfitting Problem Curves.

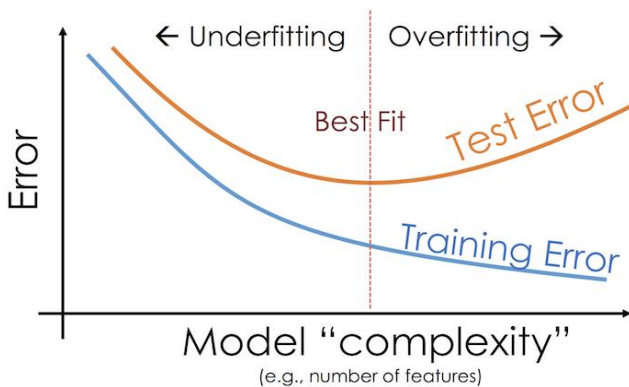


Fig. 2. The Relation between Train and Test Error and Model Order.

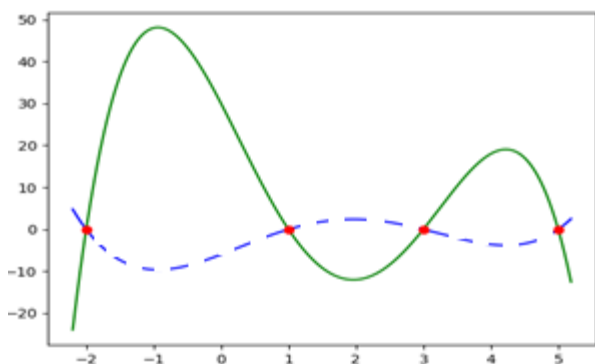


Fig. 3. Coefficients before and after Regularization.

2) Dimensionality Reduction.

3) It was hard to cope with more than a thousand features on a dataset, especially when it depended upon from where to begin! Dimensionality reduction was serving as an advantage and a defect at the same time to get a high number of variables. There was a plenty of data for the study, but the scale acted as an obstacle to get a precise information. The principle of Dimensionality reduction enabled us to to handle the extreme dimensional knowledge so that we can draw correlations and ideas from it easily. That reduction system also interfered to decrease the number of variables in the ordinary dataset by keeping a lot of data and via preserving (or improving) the model's efficiency. It was a successful attempt to operate over such huge datasets. Figure 4 illustrated that n data dimensions can be shortened into a subset of k dimensions ($k < n$).

This was called minimizing dimensionality. The advantages of dimensionality reduction on dataset were set as follows: [4]

- 1) The lower the number of measurements was, the less the space area of data storage was required.
- 2) Only Fewer measurements conform to less time for computation/training.
- 3) In the case with large dimensions, the algorithms did not accomplish well. Hence, dimensions reduction for a better performance for the algorithm was a must.
- 4) Multicollinearity was taken into consideration by excerpting superfluous features. For appetizers, two main characteristics were launched: 'time cut in minutes on the treadmill' and 'the resulted burned calories. The cause was strongly related to its effect. As the time anyone killed increasingly over a treadmill, the more calories could be burned there. Accordingly, if just one of them was done there was no need to save them.
- 5) Data visualization Support. Focusing on details in larger dimensions was not easy at all. Therefore, minimizing the distance into 2D or 3D led to more accurate areas and noticed models.

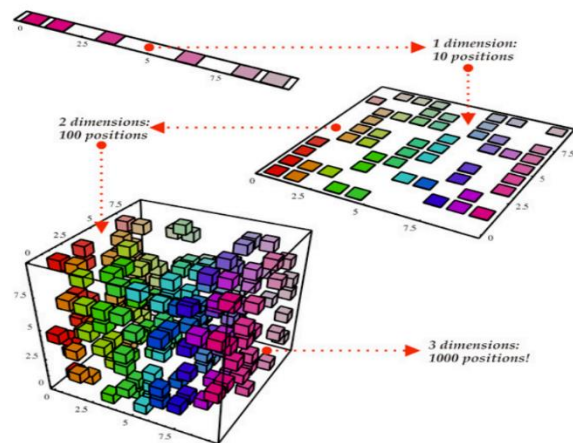


Fig. 4. Dimensionality Reduction.

The essential objective of this paper is to display how to avoid the overfitting problem for the model and improve its performance applying 9 Common dimensionality reduction techniques. The reduction technique was investigated in different dataset and their outcomes were compared through both training and testing. The evaluation was carried out depending on three separate models (ANN - SVM - RFC) with two datasets.

The remainder of discussion is arranged as follows: Section 2 contains the aimed work in brief. Section 3 tackles the description of dimensionality reduction techniques. Section 4 shows the datasets on which the experiments were taken place. It also summed up the preprocessing attitude and proved how we chose the dimensionality reduction algorithms' criterion. Section 5 traced the final results and concludes this paper.

II. RELATED WORK

The dimension reduction helped in converting data which covered a large space into a tiny space of smaller dimension [5]. In different fields, the dimension reduction was a very effective step because it facilitated the classification, visualization or compression of huge data. The purpose was to eliminate the impact of issues caused by the high dimensionality data [6].

At last, numerous methods for dimension reduction have been approached. Such ways are worthy to handle any complicated non-linear problems. This reduction was presented as a substitution to the traditional linear techniques like the PCA which is the most common way of examining nonparametric data. Once a table of measurable data (unbroken or separate) was obtained, there were no observations (individuals).

Van Der Maatan et al. [7] High-dimensional data was everywhere in computing, and all of these datasets cover a lower dimensional area than that stretched by the whole dataset. A range of dimensional reduction methods had been improved to specify this lower-dimensional space. The data map minimized the number of indicators for the supervised learning problems; besides it developed the visual performance.

Chatfield et al. [8] Key Component Analysis (PCA) and Manifold Learning are two main techniques. Nonlinear subspace mapping was the pivot of Manifold Learning and not the linear subspace mapping. Dupont et al. in [9] describes the two manners had a prominent progress in a precise description of subspace which consumed most of the data variation. Moreover, the two techniques appealed to have a clear difference. Van Der Maatan et al. [7]; a current research showed the efficacy of PCA in real datasets. The process achieved a complete success after all.

Breiman et al. [10]; the large number of dimensional reduction methods involved a variety of global, linear, non-linear, and local ways. Every attitude gathered many data characteristics. These collective attitudes led to great success in supervised learning. The idea admitted a great progress from accidental forest to KNN backsliding that aggregated to super-learners. Sollich et al. [11]; ensembles applied variety to

balance tendency, alteration, and estimator in order to carry out these effects. Therefore, those distinct techniques of dimensional reduction were probably to increase variation within a dimensional reduction ensemble. Van der Laan et al. [12] Throughout all achievements and features, it seemed that the progress of a group of local, international, linear and nonlinear dimensional reduction carriers would supply with better collective merging than any single construction. It was ordinary to differentiate the way that the superlearner ensembles. At least, it gave predictability which any model variable provided.

Dupont et al. [9] With regard to ensembles in dimensionality reduction, so tiny information had been discovered and the existing research papers did not produce ensembles with different base learner techniques. For instance, the Dupont and Ravet tried the variation of the t-distributed Stochastic Neighbor Embedding (t-SNE) criterion in their set. It was clear that it accomplished well and more effective than the tuned t-SNE model. By turning roles, it revealed the best performance among all different dimensional reduction techniques. Thus, no effort had been made to construct an ensemble using various techniques.

Zhao et al. [13] suggested a technique for dimensional reduction to the use of a spectral-space-based classification (SSFC) device to reduce the spectral dimensions. Typically, in a random way the most complicated information was taken using the Convolutional Neural Network (CNN) technique. At first, the obtained characteristics had been got and entered into the Linear Regression classifier in order to perform classification. SSFC was tested with two favorite HIS data sets, and the SVM classifier was applied to classify the images. Yan Xu et al. [14] recommended an accidental dismiss to a piece of picture from side to side with deep learning. There was a directed and unguarded structure of learning in the DNN. PCA was followed for the purpose of dimension reduction and classification. Multiple Instance Learning (MIL) had been used, too. In the beginning, natural and restricted structures should be learnt then features from the image had to be extracted from the existing dataset. The dataset consisted of high-resolution histopathological photographs of 132 patients.

The outcome conveyed that automated learning characteristics were the same like the old-fashioned set of features. Min Chen et al. [15] provided a model to help unsupervised images highlight learning for lung handling via unmarked knowledge by applying a coevolutionary, self-encoder, deep learning algorithm ;that needed a little bit of information to be called for active part learning. Autoencoder separate data information to rebuild as well as diverse input information and unique info information. Coiling autoencoder strengthened the neighborhood coiling relationship with the autoencoder to revise details for the convolution process. Dataset consisted of 4500 lung CT images from 2012 to 2015. Deep learning approaches were also investigated successfully by Yang et al. [16]. The central problems were solved in vacuum knobs examining by highlight extraction, knob detection, false-positive decline. Hence, the threatening order for the enormous volume of the father's chest filter was detected.

Deep learning also served to indicate an accurate diagnose for pneumonic knobs. The two-dimensional CNN, three-dimensional CNN, and Deep Faith Network were applied for clustering. Quantized autoencoder neural network was followed for features' derivation. An automatic technique of feature generation was presented by Rasool et al. [17] to strengthen any prediction and examine so various kinds of cancer. Unmonitored feature learning could be applied for the early detection of cancer. With the help of gene expression data, the sort of cancer could be tested. Concerning the feature learning process, softmax regression was followed as a learning procedure for the classifier. By joining 10-fold cross-validation with an aim to assess the classifier effectiveness. Hence, all the gained results were calculated showing the average accuracy of classification. A strategy for diabetic diagnosis was proposed by Y. Zheng et al. [18]. It supplied an artificial neural network. Experimental outcomes proved that this presented approach was a safe method with the situation of diabetes. It also participated in limiting the computing costs in addition to introducing accuracy. The main strategies of Highlight extraction had been planned and described to be significantly more suitable for the automatic prediction of ophthalmologist diseases than to emphasize detection methods following noisy details.

III. PRELIMINARIES

Several dimensionality reduction techniques were implemented in the following studies:

A. Missing-Values Ratio (MVR)

In (MVR), data revision should precede model construction. There was an observation that some values were missed while examining the data. An attempt was done to discover the reason behind the problem. The solution held two options whether to assign them or to remove the missing values variables at all [4]. The coming steps had to be followed respectively to get such a technique:

- Indicating the type of the missing value.
- Features determined on the ratio of missing value had to be reduced as given:

$$\text{Ratio of missing value} = \frac{\text{number of missing values}}{\text{total number of rows}} \quad (1)$$

- Rows which had a missing value such as “?, na, NULL, etc...” had to be deleted after dismissing the characteristics that had a high missing value ratio, it would be cancelled from the whole data set.

B. Low-Variance Filter (LVF)

In that method, a function was noticed to be with the same value in the dataset. With a strong observation, it wouldn't boost the formed model via this feature. Therefore, there would be zero variance in such a function.

The followed steps were stated as follows: [19]

- the variance of each feature had to be calculated.

$$\sigma^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \mu)^2 \quad (2)$$

Where, x_i stood for individual values in a dataset. μ stood for the mean of those values. n was the number of values. The term $x_i - \mu$ were named as a deviation from the mean.

- The features having low variance were dropped and compared to the most minimum value.

C. High-Correlation Filter (HCF)

It was acting as an in between technique for the ex – two. It proposed that they had specified tendency. So, similar results were expected in return. As a deduction the performance of such models would be effectively declined (e.g., linear and logistic regression models). [20] Some definite steps had to be taken to apply the (HCF) technique:

- The relation between individual numerical features had to be evaluated.
- One of the functions was completely dismissed if the correlation coefficient achieved the least value.

D. Random Forest (RFC) (for Feature Importance)

It was regarded as one of the most famous machine learning algorithms. This approach made it simple to extract each variable's value on the tree decision leaving a kind of explanation. The algorithms were so genuine because they had high detective efficiency, low overfitting, and easy interpretability.[4]. In a word, whatever each vector was joined to the decision, it was computed at once easily. As a result, a narrower subset of features [21] would be chosen.

E. Principal Component Analysis (PCA)

PCA is a linear dimensional reduction technique depended upon projection techniques. Through its application, a higher-dimensional Euclidean space had been projected into a lower-dimensional Euclidean space [4]. Given a data matrix, X , a target space, Y , and a projection matrix, P , PCA illustrated the following mapping:

$$Y = PX \quad (3)$$

The rows of P turned to a new basis for X , and by introducing this equation as an optimization problem (maximizing variance, reducing covariance between variables); it could be reconstructed in such a way that the singular value of decomposition was proposed to solve the equation. The covariance matrix, C , can be expressed as:

$$C = \frac{1}{n-1} PXX^T P^T \quad (4)$$

Because of the shift of basis found in the orthogonal bases, and because of the high variance of elements, the truncation of result would convey a mapping to a lower-dimensional space using the new bases. And at the same time a lot of variance was kept from the original dataset.

F. Linear Discriminant Analysis (LDA)

A common monitored dimensionality reduction technique was the linear discriminant analysis (LDA) [20]. LDA reached the optimal linear transformation W , which reduced the distance within the class and lengthened the distance between classes simultaneously. The criterion $J(XW)$ it maximized was:

$$J(XW) = -LDA(XW) = -\frac{W^T S_B W}{W^T S_W W} \quad (5)$$

where S_B was the between class scatter matrix and S_W was the within class scatter defined by:

$$S_B = \sum_c (\mu_c - \bar{x})(\mu_c - \bar{x})^T \quad (6)$$

$$S_W = \sum_c \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \quad (7)$$

In which \bar{x} was the mean of the data points X , μ_c and was the mean of the data points that belonged to class c .

G. Backward Feature Elimination (BFE)

To focus more and follow the 'Backward Attribute Removal' method, follow the coming steps: [22]

- The current features had to be obtained in the dataset then applied for the testing model.
- The degree of model performance had to be Calculated
- After computing the output of the model when deleting each function (n times), i.e., one variable was dropped every time and the model on the remaining n-1 variables would be tested.
- Determine the variable whose deletion got the smallest (or no) difference in the model's output, then delete that feature respectively.
- Repeat the ex-procedure many times till it was not easy for the variable to drop.

H. Forward Feature Construction (FFC)

This method was the opposite side we observed above with the Backward Attribute Removal. There was a challenge in having the right characteristics on behalf of deleting the features. Great attempts took place to reach the summit of model's performance [23]:

- The model n was tested many times starting with a single function via trying each function separately.
- The variable always provided with the best output in that indicated starting function.
- From time to time such a process was repeated through adding an element whereas the function that caused the largest progress of output was kept.
- This step was done again and again until no difference in the model efficiency was noticed.

I. Rough Set Theory (RS)

It was defined as a traditional theory created from a main research on the theoretical qualities of information systems [24]. With inaccurate and raucous data, a rough collection approach could be applied to explore any systemic relationships. The main target of this study was to activate the idea of approximation [24][25] statistical techniques were applied to display any hidden data models. Its function was to select features, derive data, and reduce details. The outlined features of reduction's fundamentals were as follows:

- The same equivalence class structure was supplied typically as that reflected by the full feature set which represented by $[x]_{RED} = [x]_P$.
- It is minimum
- It is not perfect

Algorithm 1: Reduct Calculation

Input: C, the set of all conditional features

D, the set of all decisional features

Output: R, a feature subset

1. $T := \{ \}, R := \{ \}$
 2. repeat
 3. $T := R$
 4. $\forall x \in (C - R)$
 5. if $\gamma_{RU\{X\}}(D) > \gamma_T(D)$
 6. $T := R \cup \{x\}$
 7. $R := T$
 8. until $\gamma_R(D) = \gamma_C(D)$
 9. return R
-

IV. METHODOLOGY

A. Dataset Description

The tested models were explained and confirmed with 2 classification data sets from the UCI machine-learning repository [26] Such types were involved in the experiments and comparative performance. The data sets were chosen according to various numbers of features and examples to introduce different kinds of problems on which the new approach could be examined. In addition to this, algorithm performance could be confirmed via a selection of a set of high-dimensional data. The information of training, calculating, and testing were similar in size. The training component had to be applied to train the used classifier; the validation component was used to compute the performance of the classifier; while, the evaluation component was used to evaluate the last selected characteristics which were revealed by the qualified classifier.

1) *Congressional voting records dataset:* This data set contained votes by Congressmen of the House of Representatives on the 16 primaries. Votes indicated by the CQA on each of the US. The CQA gathered nine different types of votes: voted for, paired for, and announced for (these three symbolized to yes). But, voted against, paired against, and announced against (these three simplified into no).voted present, to escape conflict of interest, and did not vote or even perform (these three simplified to an undefined provision), as in Table 1.

2) *Bands dataset:* A rotogravure printing classification query was in the shape of a cylinder unit where the aim was to define a given component. Such group of information was a UCI registry dataset, shown in Table 2.

TABLE I. CONGRESSIONAL VOTING RECORDS DATASET

Data Set Characteristics:	Multivariate	Number of Instances:	435	Area:	Social
Attribute Characteristics:	Categorical	Number of Attributes:	16	Date Donated	1987-04-27
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits	217885

TABLE II. BANDS DATASET

Data Set Characteristics:	Multivariate	Number of Instances:	512	Area:	Physical
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	39	Date Donated	1995-08-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits	77629

B. Parameters Settings

Table 3 shows parameter settings used in this study. The tested and investigated models were qualified with five hundred repetitions. ANN’s Input sheet based on No. It had thousands of hidden nodes since the helpful process was applied. This method demanded more hidden nodes than traditional algorithms. It had one output layer node introducing the 2-class. Haphazard Search CV algorithm was made to optimize the hyperparameters such like (number of estimators, max depth, ...) (the number of iterations..... ten iterations were applied). A big notice was that the greatest value of parameters was displayed in RFC. Supportive Vector Classifier plus Radial Basis Function were appeared as a Kernel. Manual tuning of hyper-parameters was served to help in choosing the support vector machine.

C. Performance Evaluation Criteria

The presented comparative models were tested depending on three variables of evaluation. Those parameters measured precision, recall, and accuracy f1 grading for both arrangement as well as research. The assessment parameters were judged as follows:

Confusion matrix:

A hesitation matrix was indicated as a table used to determine a classifier’s outcomes according to a chain of data investigation to ensure the real values (i.e., the actual positives and negatives) which were admitted.

- Precision

Precision was calculated by the ratio of results obtained via the system. It could accurately detect positive observations (True Positives) compared to the entire positive notice gained by the system, both right (True Positives) or wrong (False Positives). The accuracy equation was:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{8}$$

- Recall

The recall was the ratio of the results derived from the system-compared to all real malicious class (Actual Positives) and which in turn correctly expecting positive observations (True Positives). Hence, the recall ratio in the equation was this:

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{9}$$

- Accuracy

Precision was the most observant measure for performance. This was what many people were taught at school regardless accuracy, remember, and F1 ranking.

In short, accuracy was a change of the exacted evaluated classifications (both True Positives + True Negatives) for the whole Research Dataset. The accuracy ratio was stated:

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + false\ negatives + true\ negatives} \tag{10}$$

- F1 Score

The F1 Score was the range of Precision and Recall’s weight (or balanced mean). Consequently, to break such an equilibrium between recall and accuracy. This score had to be under focus on both false positives and false negatives. The F1 value ratio in the formula was this:

$$F1\ score = \frac{2 * (precision * recall)}{precision + recall} \tag{11}$$

TABLE III. PARAMETER SETTINGS

Model	Parameter	Values
ANN	Input nodes	based on No. Features
	Hidden nodes	1024
	Activation fun for hidden nodes	ReLU Rectified Linear Unit
	Activation fun for output nodes	sigmoid
	Output nodes	1
	No. of Iterations	500
RFC	n_estimators	1700
	max_depth	50
	min_samples_leaf	6
	class_weight	balanced
	random_state	1
SVM	n_estimators	700
	max_depth	110
	min_samples_leaf	6
	class_weight	balanced
	random_state	1

D. Model Selection

Three types of models were selected for grading (ANN – RFC - SVM). Via try/error, the most perfect hyperparameter for all models were picked out.

- As For the model of artificial neural network [27][28], its interior design mainly had Input layer, two hidden layers, and output layer.
- The input Layer based on no features.
- First hidden layer had 1024 neuron and activation mechanism was ReLU Rectified Linear Unit.
- The second hidden layer had 512 neurons, and the activation mechanism is ReLU Rectified Linear Unit.
- The output layer was 1 neuron since its class' classification problem and activation function were Sigmoid function.
- Random Forest Classifier [29].

The haphazard Search CV algorithm was applied to reach the hyperparameters to the max like (number of estimators, max depth, ...). It also supplied with the number of repetitions. Ten iterations were used, the best obtained value of parameters would be used in RFC.

- Support Vector Machine [30]

Support vector classifier had to apply the radial basis function as a Kernel. Also, a manual tuning of hyperparameters had to choose the special support vector machine.

E. Methodology and Discussion

The proposed methodology flowchart is shown in Fig. 5:

1) Dataset had to be valued a well as data preprocessing had to be applied. It was considered to be the most important step I order to create a clear ready data under usage.

2) A preprocessor with instructing data had to perform the following steps:

- a) Deleting rows which contained any missing values from dataset.
- b) Introducing the approach of "missing values ratio" algorithm.
- c) Listing categorical values.
- d) DE normalizing data (performed feature scaling).

3) If dataset couldn't be examined, only a division of 66%ran as dataset training samples, and 33% as testing ones.

4) Selection of model had to be applied.

5) Three (NN – RFC - SVM) models were selected for Classification.

6) Try/error was chosen as the best hyperparameter for all models.

7) Training and test dataset had to be proved on the selected model.

8) At last, both of detailed analysis and a comparison between results had to be applied for different models and different datasets.

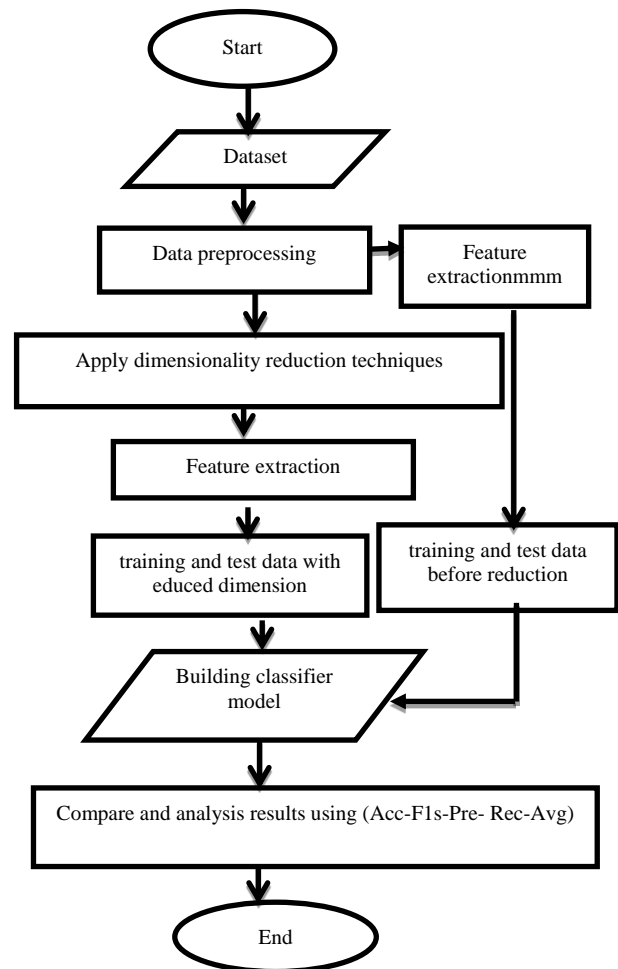


Fig. 5. Proposed Methodology Flowchart.

V. RESULTS

This quoted part showed how to find the best parameters and performance for the nine dimensionality reduction algorithms which were applied on two various datasets. It shows the bar-chart for each dataset after we have selected a comparison between the final results in a table had taken place to get the best values for the parameter in reduction methods.

A. Missing-Values Ratio

Concerning the minimum values for the ratio, the best selected one could attain the best performance. As soon as the threshold value was decreased, the number of characteristics declined with a contrast for the performance of the model which increased. By avoiding the overfitting, as revealed in Figure 6, 7, the result would be a minimum value applied on two different datasets.

B. Low-Variance Filter

With regard to this variance filter, the best minimum values would reach the best performance. As soon as the threshold value was decreased, the number of characteristics declined with a contrast for the performance of the model which increased. By avoiding the overfitting, as revealed in Figure 8, 9, the result would be a minimum value applied on two different datasets.

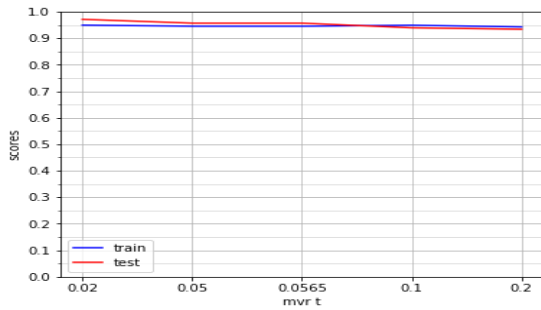


Fig. 6. Result of MVR for “Congressional Voting Records” Dataset.

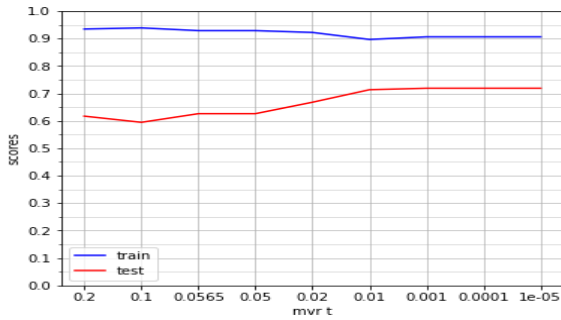


Fig. 7. Result of MVR for “Bands” Dataset

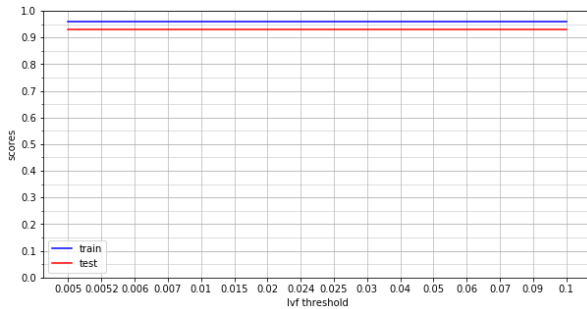


Fig. 8. Result of Low Variance Filter for “Congressional Voting Records” Dataset.

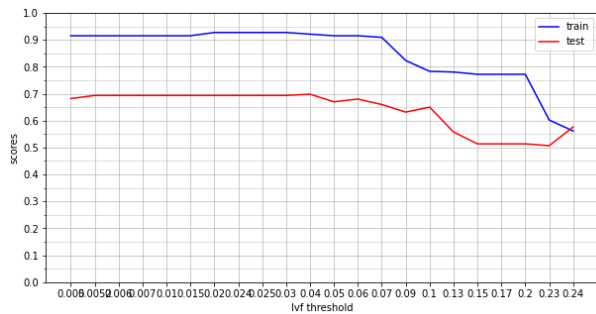


Fig. 9. Result of Low Variance Filter for “Bands” Dataset.

C. High-Correlation Filter

Following the high correlation filter, threshold values would be chosen to have the best attitude as conveyed in the figures. Whenever the threshold value was minimized, the no. features were decreasing whereas the performance of model was increasing. And this didn't allow the overfitting point to take place. Thus, no features were the same of the original

dataset. As cleared in Figure 10, 11. It provided with threshold which was applied on two different datasets.

D. Random Forest

In a random forest, an enormous accurate built chain of trees was created against achieving the highest features. Therefore, the usage statistics of each characteristic was used to obtain the greatest instructive subset of feature s. Figure 12 and 13 illustrated threshold which was applied on two different datasets.

E. Principal Component Analysis

Minimizing No. Features/Dimensions and applying PCA on them in order to choose the best value for PCA. From ex-knowledge, no. PCA was probably from 1: n-1 features. Figure 14 and 15 applied such a minimization on two different datasets.

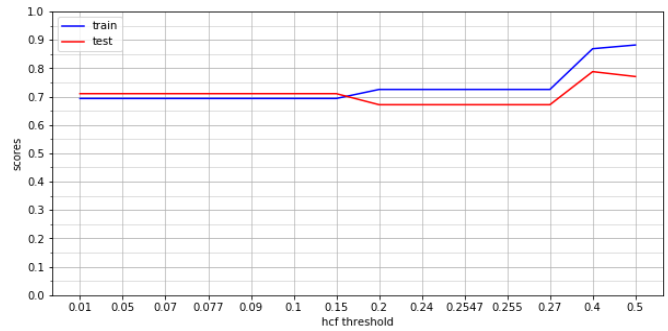


Fig. 10. Result of High-Correlation Filter for “Congressional Voting Records” Dataset.

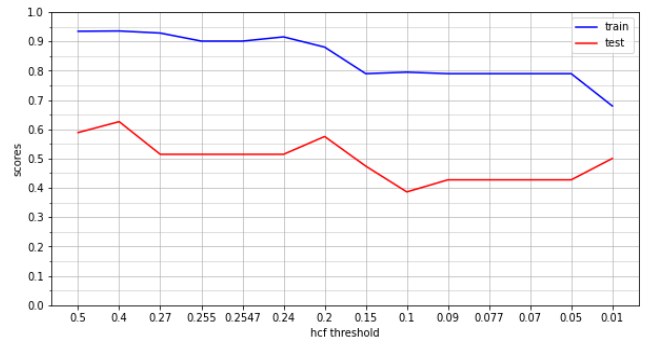


Fig. 11. Result of High-Correlation Filter for “Bands” Dataset.

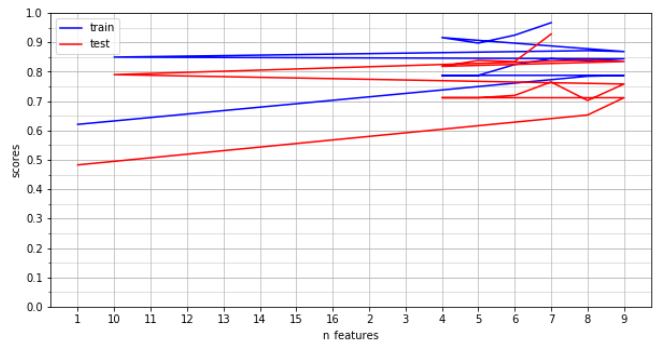


Fig. 12. Result of Random Forest for “Congressional Voting Records” Dataset.

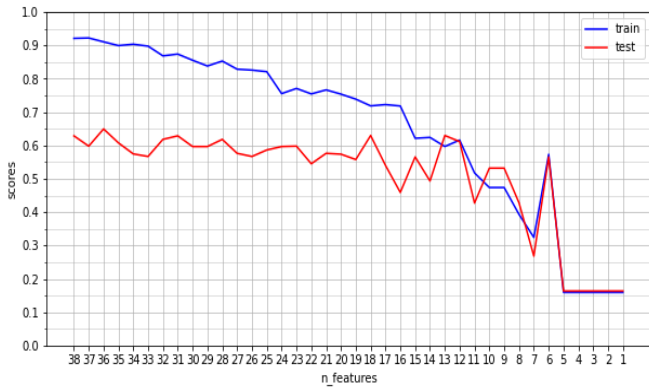


Fig. 13. Result of Random Forest for "Bands" Dataset.

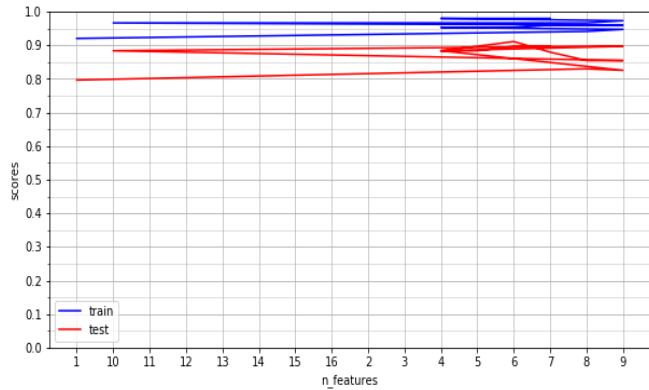


Fig. 14. Result of Principal Component Analysis for "Congressional Voting Records" Dataset.

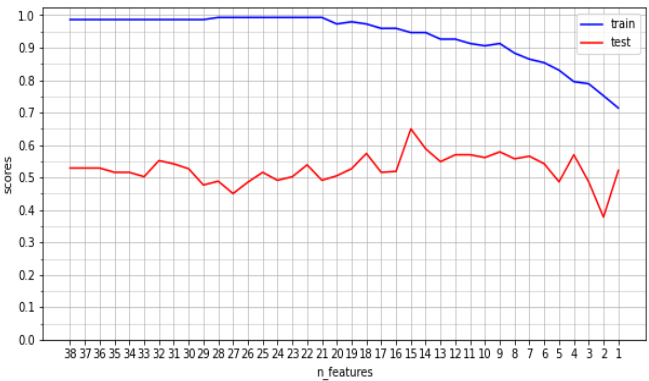


Fig. 15. Result of Principal Component Analysis for "Bands" Dataset.

F. Linear Discriminant Analysis

LDA was carried out to get fitting training data by giving new sized area since its reduction technique built on maximizing the class severability. If there were two classes No. LDA will be 1. The minimum value was applied on two different datasets as described in figure 16 and 17.

G. Backward Feature Elimination

Gradual decline of No Features with a threshold value to reach the best features and threshold values that achieve the best performance. As shown in Figure 18 and 19 applied on 2 different datasets.

H. Forward Feature Construction

The definition of forward feature construction was the opposite of the backward technique. It occurred by raising No. Features with a threshold value to have the best characteristics and the threshold value which achieved the perfect performance. As shown in Figure 20 and 21, it was applied on two different datasets.

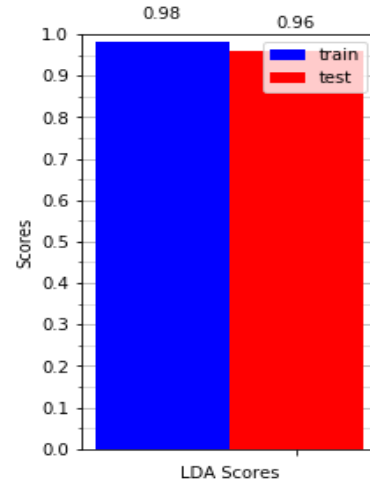


Fig. 16. Result of LDA for "Congressional Voting Records" Dataset.

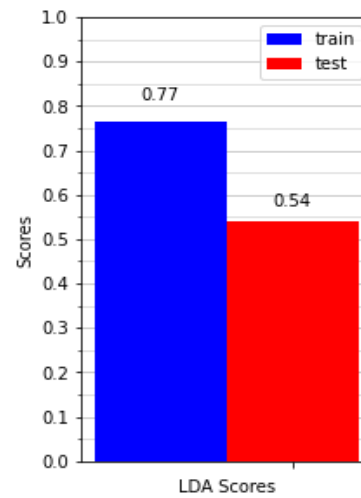


Fig. 17. Result of Linear Discriminant Analysis for "Bands" Dataset.

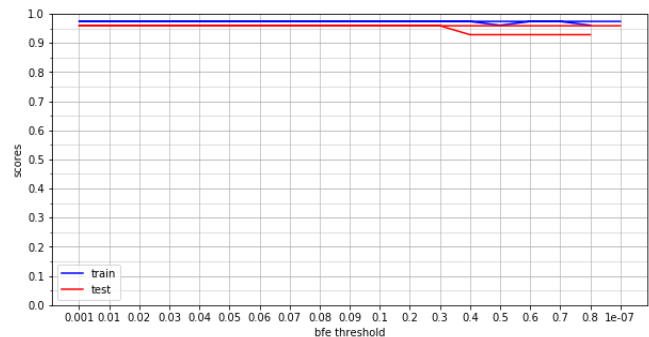


Fig. 18. Result of Backward Feature Elimination for "Congressional Voting Records" Dataset.

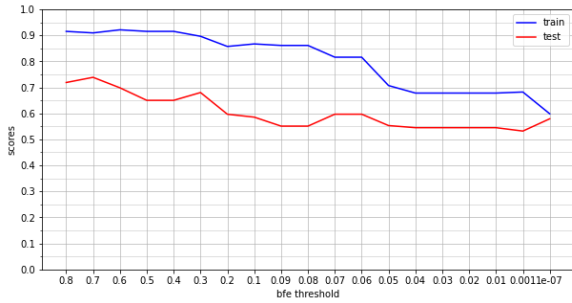


Fig. 19. Result of Backward Feature Elimination for “Bands” Dataset.

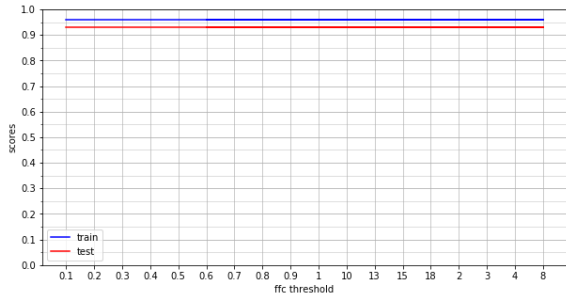


Fig. 20. Result of Forwarding Feature Construction For “Congressional Voting Records” Dataset.

I. For “Congressional Voting Records” Dataset

The detailed analysis of tables 4 and 5, the unique reduction techniques were Low-Variance Filter (LVF) in the training dataset and Missing-values Ration (MVR) in the test dataset. In tables six, the greatest reduction approach was Linear Discriminant Analysis (LDA) in both training & testing dataset. Performance, here, was so near compared to the actual overfitting which was slightly reduced (it’s challenging to minimize overfitting without losing preciseness because the dataset was very tiny.

The outcomes of all techniques with average results as (avg) from three models on training & testing dataset (Congressional Voting Records Dataset). The best reduction

method was applied on the Missing-Values Ratio symbolized (MVR) as shown in Figure 22.

- The number of features was reduced from 16 to 10
- Reduction percentage = 62.5%
- Performance was improved by 3% (from 94% to 97%)
- Overfitting was decreased by 2%

The result of the test score for the three models for each reduction technique (Congressional Voting Records Dataset) for the Best model was NN as described in Figure 23.

J. For “Bands” Dataset

In Tables 7 and 8, the best reduction method was Principal Component Analysis (PCA) in the training dataset and Random Forest (RFC) in the testing dataset. But for Table 9, the best reduction method was the Principal Component Analysis (PCA) in the training dataset and Missing-Values Ratio (MVR) in the testing dataset. At the same time performance was improved and overfitting was slightly decreased.

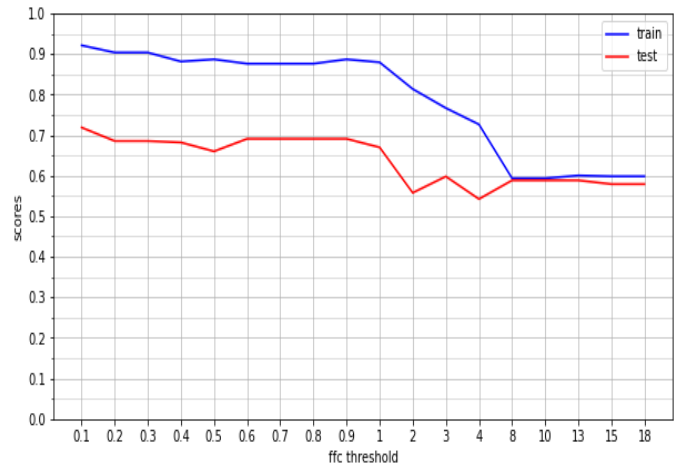


Fig. 21. Result of Forwarding Feature Construction for “Bands” Dataset.

TABLE IV. RESULTS OF ALL TECHNIQUES WITH ANN MODEL ON TRAINING & TESTING DATASET

		ANN									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	1.00	1.00	1.00	1.00	1.00	0.96	0.96	1.00	0.92	0.96
2	Missing-Values Ratio	0.98	0.98	0.99	0.96	0.98	0.99	0.99	1.00	0.98	0.99
3	Low-Variance Filter	0.99	0.99	1.00	0.99	0.99	0.96	0.96	1.00	0.92	0.96
4	High-Correlation Filter	0.67	0.63	0.57	0.69	0.64	0.65	0.63	0.73	0.56	0.64
5	Random Forest	0.90	0.89	0.83	0.95	0.89	0.85	0.80	0.73	0.89	0.82
6	Principal Component Analysis	0.98	0.98	0.97	0.99	0.98	0.95	0.94	0.97	0.91	0.94
7	Linear Discriminant Analysis	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.98	0.98	0.99	0.97	0.98	0.95	0.94	0.94	0.94	0.94
10	Rough-set	0.94	0.92	0.92	0.92	0.92	0.89	0.86	0.90	0.82	0.87

TABLE V. RESULT OF ALL TECHNIQUES WITH RFC MODEL ON TRAINING & TEST DATASET

		RFC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
2	Missing-Values Ratio	0.95	0.94	0.97	0.92	0.95	0.97	0.95	0.98	0.93	0.96
3	Low-Variance Filter	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
4	High-Correlation Filter	0.69	0.73	0.85	0.63	0.73	0.62	0.66	0.88	0.53	0.67
5	Random Forest	0.92	0.92	0.95	0.89	0.92	0.84	0.81	0.85	0.78	0.82
6	Principal Component Analysis	0.98	0.98	0.97	0.99	0.98	0.91	0.89	0.88	0.91	0.90
7	Linear Discriminant Analysis	0.98	0.98	0.99	0.97	0.98	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.96	0.96	0.97	0.95	0.96	0.94	0.93	0.94	0.91	0.93
10	Rough-set	0.93	0.91	0.93	0.88	0.91	0.87	0.84	0.90	0.79	0.85

TABLE VI. RESULT OF ALL TECHNIQUES WITH SVC MODEL ON TRAINING & TEST DATASET

		SVC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
2	Missing-Values Ratio	0.98	0.97	0.98	0.96	0.97	0.97	0.97	1.00	0.93	0.97
3	Low-Variance Filter	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
4	High-Correlation Filter	0.69	0.73	0.85	0.63	0.73	0.62	0.66	0.88	0.53	0.67
5	Random Forest	0.95	0.95	0.97	0.94	0.95	0.82	0.78	0.76	0.81	0.79
6	Principal Component Analysis	0.98	0.98	0.99	0.97	0.98	0.94	0.93	0.94	0.91	0.93
7	Linear Discriminant Analysis	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
8	Backward Feature Elimination	0.97	0.97	0.99	0.96	0.97	0.96	0.96	1.00	0.92	0.96
9	Forward Feature Construction	0.97	0.97	0.97	0.97	0.97	0.94	0.93	0.94	0.91	0.93
10	Rough-set	0.96	0.95	0.97	0.94	0.96	0.89	0.86	0.92	0.81	0.87

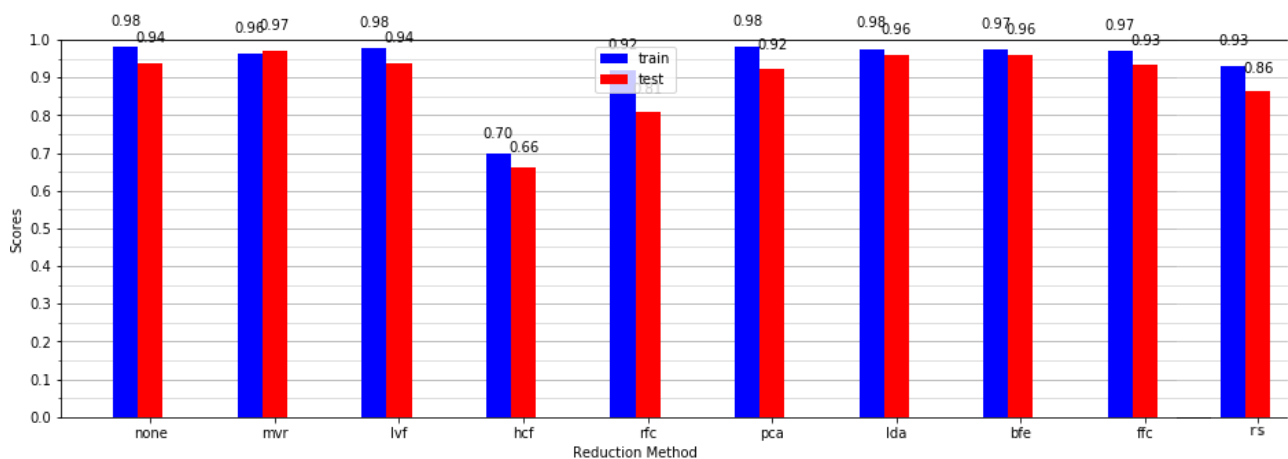


Fig. 22. Result of All Techniques with Average Results of (AVG) from Three Models on Training & Testing Dataset.

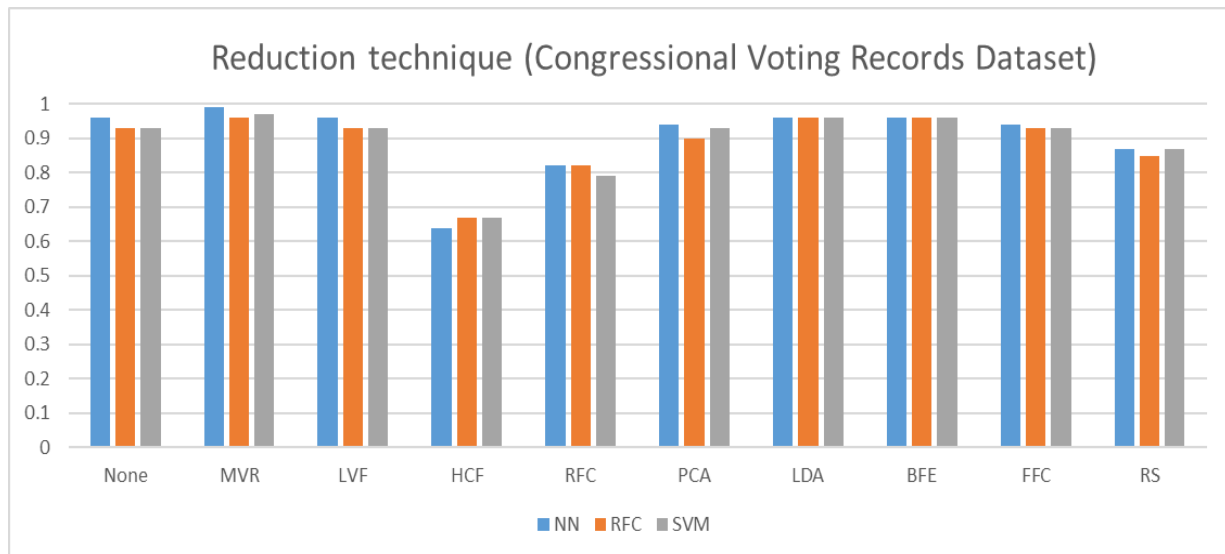


Fig. 23. Test Score for Three Models for each Reduction Technique (Congressional Voting Records Dataset).

TABLE VII. RESULT OF ALL TECHNIQUES WITH THE ANN MODEL ON TRAINING & TESTING DATASET

		ANN									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.74	0.72	0.94	0.58	0.75	0.64	0.56	0.66	0.49	0.59
2	Missing-Values Ratio	0.65	0.70	0.97	0.55	0.72	0.51	0.62	0.97	0.46	0.64
3	Low-Variance Filter	0.64	0.65	0.90	0.50	0.67	0.56	0.51	0.66	0.41	0.53
4	High-Correlation Filter	0.67	0.58	0.61	0.54	0.60	0.54	0.26	0.24	0.29	0.33
5	Random Forest	0.63	0.62	0.81	0.50	0.64	0.68	0.58	0.66	0.53	0.61
6	Principal Component Analysis	0.76	0.73	0.91	0.61	0.75	0.67	0.53	0.55	0.52	0.57
7	Linear Discriminant Analysis	0.76	0.69	0.74	0.65	0.71	0.67	0.50	0.48	0.52	0.54
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.60	0.59	0.80	0.47	0.61	0.52	0.47	0.62	0.38	0.50
10	Rough-set	0.62	0.41	0.37	0.48	0.47	0.55	0.36	0.37	0.35	0.41

TABLE VIII. RESULT OF ALL TECHNIQUES WITH RFC MODEL ON TRAINING & TESTING DATASET

		RFC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.89	0.86	0.94	0.79	0.87	0.69	0.55	0.55	0.55	0.59
2	Missing-Values Ratio	0.78	0.76	0.82	0.71	0.77	0.66	0.66	0.80	0.56	0.67
3	Low-Variance Filter	0.73	0.70	0.89	0.58	0.72	0.63	0.55	0.66	0.47	0.58
4	High-Correlation Filter	0.74	0.66	0.70	0.63	0.68	0.56	0.24	0.21	0.30	0.33
5	Random Forest	0.67	0.64	0.83	0.53	0.67	0.67	0.58	0.66	0.51	0.60
6	Principal Component Analysis	0.84	0.81	0.93	0.72	0.83	0.67	0.53	0.55	0.52	0.57
7	Linear Discriminant Analysis	0.76	0.68	0.73	0.65	0.70	0.67	0.48	0.45	0.52	0.53
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.64	0.58	0.69	0.50	0.60	0.68	0.56	0.59	0.53	0.59
10	Rough-set	0.69	0.70	0.63	0.79	0.70	0.56	0.60	0.55	0.66	0.59

TABLE IX. RESULT OF ALL TECHNIQUES WITH SVC MODEL ON TRAINING & TEST DATASET

		SVC									
		Train Dataset					Test Dataset				
		Acc	F1s	Pre	Rec	Avg	Acc	F1s	Pre	Rec	Avg
1	Before Reduction	0.94	0.92	0.93	0.90	0.92	0.75	0.63	0.62	0.64	0.66
2	Missing-Values Ratio	0.92	0.90	0.90	0.90	0.91	0.74	0.71	0.76	0.67	0.72
3	Low-Variance Filter	0.81	0.77	0.87	0.69	0.78	0.71	0.62	0.69	0.57	0.65
4	High-Correlation Filter	0.91	0.87	0.87	0.87	0.88	0.71	0.52	0.45	0.45	0.58
5	Random Forest	0.63	0.57	0.69	0.49	0.60	0.70	0.60	0.66	0.56	0.63
6	Principal Component Analysis	0.96	0.94	0.94	0.94	0.95	0.74	0.62	0.62	0.62	0.65
7	Linear Discriminant Analysis	0.81	0.75	0.77	0.73	0.73	0.68	0.49	0.45	0.54	0.54
8	Backward Feature Elimination	0.63	0.58	0.70	0.49	0.60	0.67	0.55	0.59	0.52	0.58
9	Forward Feature Construction	0.63	0.57	0.69	0.49	0.59	0.68	0.56	0.59	0.53	0.59
10	Rough-set	0.86	0.87	0.84	0.91	0.87	0.68	0.50	0.47	0.53	0.55

The result of all techniques with average results of (avg) from three models on training and testing dataset (Bands Dataset), applying the best reduction method, was Missing-Values Ratio (MVR) as reflected in Figure 24.

- The number of features reduced from 38 to 12
- Reduction percentage = 68%
- Performance was improved by 7% (from 66% to 72%)
- Overfitting was reduced

The result of the test score for the 3 models for each reduction technique (Bands Dataset) applying the best model was RFC as shown in Figure 25 reduction techniques. Those techniques were called as follows: missing-values ratio, low variance filter, high correlation filter, random forest, key component analysis, linear discriminant analysis, removal of backward function, construction of forwarding features, and rough set theory. It was observed how dimensionality reduction could be useful in minimizing overfitting as well as getting the perfect performance. The models' average performance was tested in two various datasets for three different models, as cleared in Table 10.

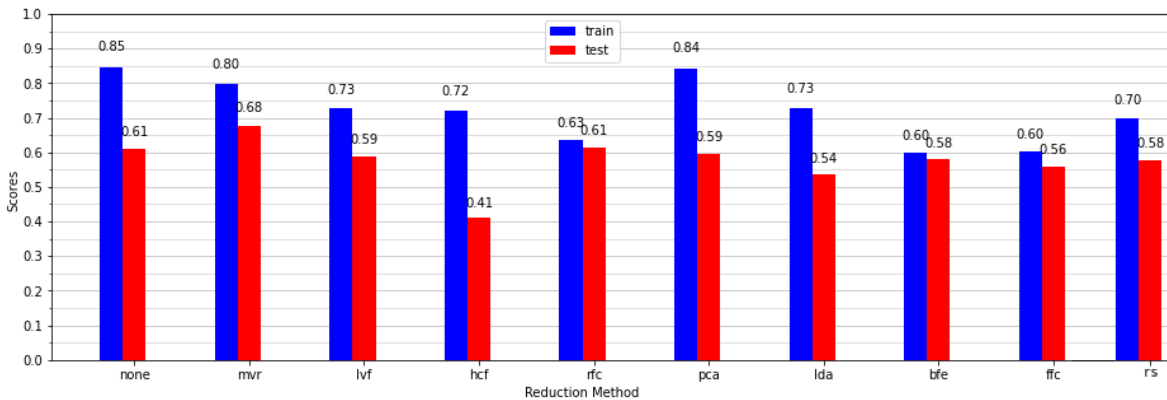


Fig. 24. Result of All Techniques with Average Results of (AVG) from Three Models on Training & Testing.

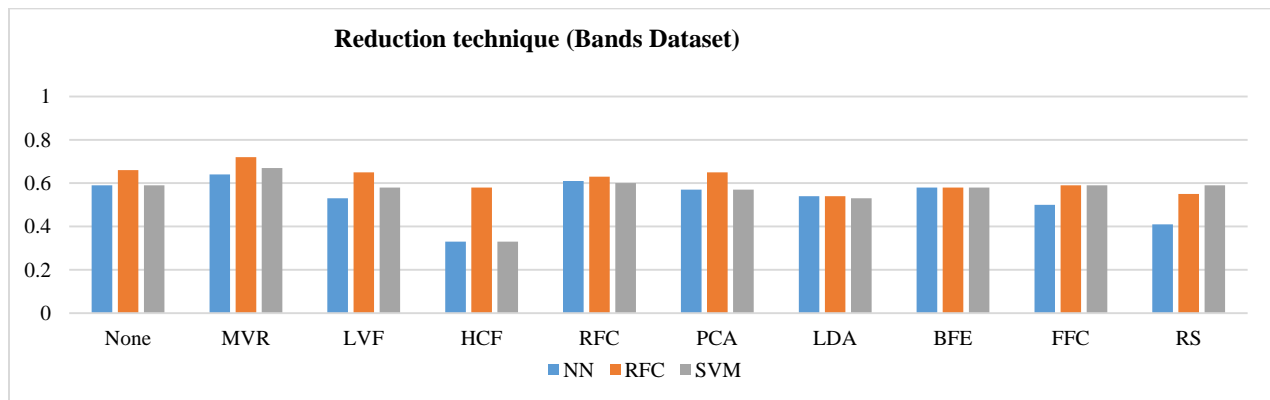


Fig. 25. The Test Score of Three Models for each Reduction Technique (Bands Dataset) Described a Number of Nine-Dimensional.

TABLE X. PERFORMANCE EVALUATION OF DIFFERENT MODELS ON TEST DATASET

		Congressional Voting Records Database			Bands Database		
		ANN	RFC	SVC	ANN	RFC	SVC
		Avg	Avg	Avg	Avg	Avg	Avg
1	Before Reduction	0.96	0.93	0.93	0.59	0.59	0.66
2	Missing-Values Ratio	0.99	0.96	0.97	0.64	0.67	0.72
3	Low-Variance Filter	0.96	0.93	0.93	0.53	0.58	0.65
4	High-Correlation Filter	0.64	0.67	0.67	0.33	0.33	0.58
5	Random Forest	0.82	0.82	0.79	0.61	0.60	0.63
6	Principal Component Analysis	0.94	0.90	0.93	0.57	0.57	0.65
7	Linear Discriminant Analysis	0.96	0.96	0.96	0.54	0.53	0.54
8	Backward Feature Elimination	0.96	0.96	0.96	0.58	0.58	0.58
9	Forward Feature Construction	0.94	0.93	0.93	0.50	0.59	0.59
10	Rough-set	0.87	0.85	0.87	0.41	0.59	0.55

VI. CONCLUSION

This paper discussed nine-dimensional reduction techniques, and their effect on overfitting problem. These techniques are namely, missing-values ratio, low variance filter, high correlation filter, random forest, key component analysis, linear discriminant analysis, removal of backward function, construction of forwarding features, and rough set theory respectively. These techniques are valuable in reducing overfitting as well as obtaining a quite accepted performance. The used techniques were compared in both training and testing performance on two different datasets with three different models (ANN, SVM, and RFC). Performance was so close to the original for the RFC model. Missing-values ratio was closer to the removal of backward feature. The datasets got reduced to almost half of their original size; that allows machine-learning models to work faster on the datasets, which was another advantage of dimensionality reduction.

Some improvements on used models will be added by using metaheuristic optimization algorithms to find the best solution.

REFERENCES

[1] C. L. Blake, "CJ Merz UCI repository of machine learning databases," Ph.D. dissertations, Dept. Inform. Comput. Sci., Univ. California, Irvine, CA, USA, 1998.

[2] O. Deniz, A. Pedraza, N. Vallez, J. Salido, and G. Bueno, "Robustness to adversarial examples can be improved with overfitting". International Journal of Machine Learning and Cybernetics, 1-10, 2020.

[3] A. Zheng and A. Casari, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists, Newton, MA, USA:O'Reilly Media, 2018.

[4] X. Huang, L. Wu, and Y. Ye, "A Review on Dimensionality Reduction Techniques," International Journal of Pattern Recognition and Artificial Intelligence, vol. 33, no. 10, p. 1950017, 2019.

[5] A. Juvonen, T. Sipola, T. Hämmäläinen, Online anomaly detection using dimensionality reduction techniques for http log analysis, Comput. Netw. 91 (2015) 46–56.

[6] M. Verleysen, D. François, The Curse of Dimensionality in Data Mining and Time Series Prediction, in: International Work-Conference on Artificial Neural Networks, Springer, 2005, pp. 758–770.

[7] T. Lesort , N. Díaz-Rodríguez , J.-F. Goudou , D. Filliat , State representation learning for control: an overview, Neural Netw. 108 (2018) 379–392 .

[8] C. Meng, O.A. Zeleznik, G.G. Thallinger, B. Kuster, A.M. Gholami, A.C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data, Brief. Bioinform. 17 (4) (2016) 628–641.

[9] J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: survey, insights, and generalizations, J. Mach. Learn. Res. 16 (1) (2015) 2859–2900.

[10] L. Xie, Z. Li, J. Zeng, U. Kruger , Block adaptive kernel principal component analysis for nonlinear process monitoring, AIChE J. 62 (12) (2016) 4334–4345.

[11] A. Akkalkotkar, K.S. Brown, An algorithm for separation of mixed sparse and gaussian sources, PloS one 12 (4) (2017) e0175775.

[12] S. Deegalla , H. Boström , K. Walgama , Choice of Dimensionality Reduction Methods for Feature and Classifier Fusion with Nearest Neighbor Classifiers, in: 15th International Conference on Information Fusion (FUSION), IEEE, 2012, pp. 875–881.

[13] S. Ahmadkhani, P. Adibi, Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework, IET Comput. Vision 10 (3) (2016) 193–201.

[14] I.T. Jolliffe , J. Cadima , Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202.

[15] NB. Erichson, P. Zheng, K. Manohar, S.L. Brunton, J.N. Kutz, A.Y. Aravkin, Sparse principal component analysis via variable projection. arXiv preprint arXiv:1804.00341.

[16] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," Computers & Electrical Engineering, vol. 70, pp. 871-882, 2018.

[17] S. Chormunge and S. Jena, "Correlation-based feature selection with clustering for high dimensional data," Journal of Electrical Systems and Information Technology, vol. 5, no. 3, pp. 542-549, 2018.

[18] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," in 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 2264-2268: IEEE.

[19] B. Liu, Y. Li, L. Li, and Y. Yu, "An approximate reduction algorithm based on conditional entropy." In International Conference on Information Computing and Applications (pp. 319-325). Springer, Berlin, Heidelberg, 2010.

[20] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. A. Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," Computers & Electrical Engineering, vol. 70, pp. 871-882, 2018.

[21] S. Hu , Y. Gu , H. Jiang , Study of Classification Model for College Students' M-learn- ing Strategies Based on Pca-lvq Neural Network, in:

- 8th International Conference on Biomedical Engineering and Informatics (BMEI), IEEE, 2015, pp. 742–746.
- [22] A.R. Santos, M.A. Santos, J. Baumbach, J.A.M. Culloch, G.C. Oliveira, A. Silva, A. Miyoshi, V. Azevedo, A Singular Value Decomposition Approach for Improved Taxonomic Classification of Biological Sequences, in: BMC Genomics, volume 12, BioMed Central, 2011, p. S11.
- [23] A. Swati, and R. Ade, “Dimensionality reduction: an effective technique for feature selection”. *Int. J. Comput. Appl.* 117(3), 18–23, 2015.
- [24] S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” *Information Fusion*, vol. 59, pp. 44-58, 2020.
- [25] P. Mills, Singular value decomposition (svd) tutorial: Applications, examples, exercises, 2017, <https://blog.statsbot.co/singular-value-decomposition-tutorial-52c695315254>, (Accessed on 09/04/2019).
- [26] I. Brigadir, D. Greene, J. P. Cross, and P. Cunningham, “Dimensionality Reduction and Visualisation Tools for Voting Record.” In 24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS’16), University College Dublin, Ireland, 20-21 September 2016. CEUR Workshop Proceedings
- [27] S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, “Applications of support vector machine (SVM) learning in cancer genomics.” *Cancer Genomics-Proteomics*, 15(1), 41-51, 2018.
- [28] J. Rahmanishamsi, A. Dolati, M.R. Aghabozorgi, A copula based algorithm and its application to time series clustering, *J. Classif.* 35 (2) (2018) 230–249.
- [29] Q. Hu, L. Zhang, Y. Zhou & W. Pedrycz, “Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets.” *IEEE Transactions on Fuzzy Systems* (1), 226-238, 2018.
- [30] (1), 226-238, 2018.
- [31] A. Zeng, T. Li, D. Liu, J. Zhanga and H. Chen, “A fuzzy rough set approach for incremental feature selection on hybrid information systems” *Fuzzy Sets and Systems*, 258, 39-60, 2015.

A Sentiment Analysis of Egypt's New Real Estate Registration Law on Facebook

Abdulfattah Omar¹

Department of English
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said
University, Egypt

Wafya Ibrahim Hamouda²

Department of Foreign Languages
Faculty of Education
Tanta University
Tanta, Egypt

Abstract—In response to the increasing influence of social media networks on shaping the public opinion, sentiment analysis systems and applications have been developed to extract insights and gain an overview of the wider public opinion behind certain topics so as to support businesses, manufacturers, government agencies, and policymakers with their decisions and plans. Despite the importance of sentiment analysis in providing policymakers with effective mechanisms to understand the attitudes of customers and citizens which can be usefully used in decision-making processes and planning for the future, so far studies on sentiment analysis are very limited in Egypt. Much of the work is still done using survey tools such as questionnaires and polls to gather information about the citizens' attitudes towards given issues and topics. Despite the effectiveness of such methods, citizens' reflections on social media platforms and networks remain more powerful in providing comprehensive insights and overviews. Furthermore, social media-based sentiment analysis is usually more representative being based on larger numbers of participants, which has positive implications to reliability. Opinions expressed on social media are often the most powerful forms of feedback for businesses because they are given unsolicited. In light of this argument, this study seeks to provide a sentiment analysis of Egypt's New Real Estate Registration Law on Facebook. To extract information about the users' sentiment polarity (positive, neutral or negative), Facebook posts were used. The rationale is that Facebook is still the most popular social media platform in Egypt. Text classification was then used for classifying the selected data into three main classes/values: Positive, Negative, and Neutral. The findings indicate that sentiments expressed in the users' posts and comments have a significant negative attitude towards the new law. Despite the effectiveness of the automatic evaluation and analysis of the sentiments and opinions of the users in social media concerning the new Real Estate Registration Law, linguistic approaches including Critical Discourse Analysis (CDA), functional linguistics, and semiotics need to be incorporated into sentiment analysis applications for gaining a better understanding of people's attitudes towards specific issues.

Keywords—Egypt; Facebook; opinion; real estate registration law; sentiment analysis; social media

I. INTRODUCTION

The role of social media platforms and networks has grown drastically over the recent years. Today, social media platforms and networks are not only accessible and easy venues and channels for communication between individuals

across the globe, they are among the most important tools for influencing the public opinion and decision making processes [1-3]. This is obviously reflected in the influential role social media played in the social and political movements in different countries in the world [4, 5].

In the Arab world, social media platforms and networks played a vital role during the so-called 'Arab Spring events' in political mobilization by calling for demonstrations and protests, publishing news and videos, expressing opinions, and political debate liberated from the authority's control over the traditional media [6]. This role was definitely supported by the potentials of social media of reaching millions of users through very high interactive features within a very short period of time and forming discussion forums with large numbers of participants.

In recent years, there are many examples where social media influenced public opinion through forming a unified public opinion on specific issues as a result of the interaction between its users who belong to different cultures, but who believe in a common system of values. In 2014, for instance, the Canadians expressed their anger against Expedia Canada television advertisement on the social media networks especially Facebook and Twitter. The "Escape Winter: Fear" advertisement caused a stir on social media when the frequency of which it aired proved to be too much for some viewers [7]. Viewers from different social and cultural backgrounds placed tormented posts on Facebook and Twitter which was described as a public negative attitude towards the company.

Given the increasing importance of social media in public opinion, sentiment analysis has been developed in order to extract insights and gain an overview of the wider public opinion behind certain topics. In marketing, sentiment analysis is widely used to help brands, businesses and manufacturers address and handle customer anger, protect the brand image, and evaluate the effectiveness of their marketing campaigns [8-10]. In this sense, sentiment analysis has been one of the most common and fundamental measures of customers' attitudes toward a brand by using different variables including language constructs, emotion, and so on. Manufacturers, businesses, and marketing agencies use sentiment analysis to learn more about how people feel about their business and goods, as well as to assess customer loyalty [11-14].

Recently, sentiment analysis has been widely used by government agencies and policy makers. Prior to the 2012 presidential election, for instance, the Obama administration used opinion analysis to gauge public response to policy announcements and campaign messages. The underlying principle is that understanding and interpreting citizens' concerns in real-time can be usefully used in improving public services, building well-organized governments, and creating predictive models to anticipate bottlenecks in various public services [15, 16]. According to Arunachalam and Sarkar [17], sentiment analysis of the citizens' posts on social media networks is best considered as the new eye of the government.

Despite the importance of sentiment analysis in providing policymakers with effective mechanisms to understand the attitudes of customers and citizens which can be usefully used in decision-making processes and planning for the future, so far studies on sentiment analysis are very limited in Egypt. Much of the work is still done using survey tools such as questionnaires and polls to gather information about the citizens' attitudes towards given issues and topics. Despite the effectiveness of such methods, citizens' reflections on social media platforms and networks remain more powerful in providing comprehensive insights and overviews. Furthermore, social media-based sentiment analysis is usually more representative being based on larger numbers of participants, which has positive implications to reliability. Opinions expressed on social media are often the most powerful forms of feedback for businesses because they are given unsolicited. In light of this argument, this study seeks to provide a sentiment analysis of Egypt's New Real Estate Registration Law on Facebook.

In August 2020, the House of Representatives in Egypt (the Egyptian Parliament) approved a draft law submitted by the Government regarding the amendment of some provisions of Law No. 114 of 1946 regulating the real estate registry in all of its articles, by adding an updated Article No. 35 (bis) in order to protect the real estate wealth and its rights, according to the government officials [18]. However, the amendments meant stopping the introduction of utilities for all unregistered buildings as of March 6, 2021. By the law, electricity, water and gas companies and other companies, agencies, Ministries and government departments are not allowed to transfer facilities and services to properties unless they are officially registered in one of the registry offices. This amendment was published in the Official Gazette on September 5, 2020, to come into effect in six months, i.e. on March 6, 2021.

On September 9, 2020 the Prime Minister indicated that the government is currently working on setting up a digital system to create a certificate and a national number for each apartment and property, and apartments will only be dealt with. On February 22, 2021, the Minister of Justice confirmed that all government agencies will not deal with unregistered real estate, meaning that facilities will not be connected to such unregistered units starting from March 6, 2021 [19].

As a result, there was a lot of discussion, controversy, and uncertainty on social media networks in Egypt. Some considered the issuance of the law in its new form an achievement that makes each apartment and real estate a

national number, and accurately determines the size of the country's real estate wealth, as well as gives real estate an actual and substantial value. Others called on the government to review the controversial law and simplify the procedures for registering housing units [20].

This paper discusses the controversy that took place on social media networks with the enforcement of the new amendment on March 6, 2021. It seeks to address the following research questions:

1) What were the sentiments, opinions, and attitudes of the Egyptians towards the new Real Estate Registration Law as expressed on their posts, comments, and replies on Facebook?

2) What is the effectiveness of automated sentiment analysis in evaluating the opinions and responses of citizens on social media towards regulations and government policies as represented in the analysis of the sentiments, opinions, and attitudes of the Egyptians towards the new Real Estate Registration Law as expressed on their posts, comments, and replies on Facebook?

The rest of the paper is organized as follows. Section 2 is an introduction to the concept of sentiment analysis. Section 3 is a brief survey of the development of sentiment analysis systems and applications in Arabic. Section 4 describes the methods and procedures of data collection, preparation, and classification. Section 5 reports the results of the study. Section 6 is conclusion.

II. SENTIMENT ANALYSIS

Sentiment analysis is a relatively new discipline of knowledge. It emerged around the year 2000 with the development of social networks and platforms. Nevertheless, it has been very active since its inception. Sentiment analysis has been associated with different disciplines including behavior analysis, business administration, data mining, information retrieval, marketing, Natural Language Processing (NLP), text classification, and text mining, and web mining [21-24].

Apparently, the recent years have witnessed an accelerating development of sentiment analysis due to the explosive growth of the web and the advent of social media platforms and networks over the past twenty years. Sentiment analysis, also called opinion mining, Pozzi, et al. [21] argue, has come to be extensively used to evaluate people's opinions, sentiments, appraisals, judgments, attitudes, and emotions toward products, services, organizations, individuals, events, issues, or topics as expressed in written text, emotions, and so on through their posts and comments on social media platforms and networks.

Iglesias and Moreno [22] assert although sentiment analysis can be applied to different data sources including surveys, opinion polls, and focus group discussions, the advent of social media systems has given sentiment analysis more value and added to its popularity. In other words, social media networks, platforms, and systems have provided sentiment analysts with rare opportunities to construct

organized and actionable knowledge through defining automated tools that can extract subjective information [25]. In other words, the prolific data on the social media networks, platforms, and systems have made it possible for organizations to automate the analysis of opinions expressed in digital form [24].

In recent years, different sentiment analysis systems have been developed to score and quantify the sentiments towards issues, goods, services, and so on. These have been developed using different software languages including R and Python [26]. In these model, the objective is to identify emotions, such as: happiness, fear, anger, etc. Glossaries are usually generated using machine learning algorithms to define these feelings. To put it into effect, these are lists of words with corresponding emotions associated with them. Here it should be noted that when using these glossaries, the problem of different emotions conveyed by words appears, especially that humans can express their emotions in different ways.

III. RELATED WORK

Arabic sentiment analysis was first introduced in 2008. Progress on Arabic sentiment analysis research was very slow. During this early period which extended until approximately 2015, very few studies were done. This can be attributed to the idea that the main bulk of research in sentiment analysis was traditionally focused on English. Very few studies were done on other languages due to the lack of resources that support the analysis of sentiments in other domains [27, 28].

During this early or first stage, Arabic sentiment analysis studies were concerned with developing sentiment analysis systems taking into consideration the peculiar linguistic features of Arabic. In order to support sentiment analysis applications in Arabic, Rushdi-Saleh, et al. [27] built a corpus of 500 movie reviews collected from different web pages and blogs in Arabic. Similarly, Abdul-Mageed, et al. [29] attempted to address the linguistic challenges that are always associated with sentiment analysis through developing a sentiment analysis system based on sentence-level that takes into consideration the morphological system of Arabic. The proposed model was able to achieve higher levels of performance.

To bridge the gap between sentiment analysis theory and applications, Elhawary and Elfeky [30] developed an Arabic sentiment analysis system that enables businesses and corporations in the Middle East to better understand the experiences of their Arab customers in the region. The proposed system was largely based on lexicon-based methods and it showed good performance compared to the English sentiment analysis. Despite the relative success of these studies during this early stage to introduce sentiment analysis in Arabic for the first time, research and applications in Arabic sentiment analysis were very limited.

The rise of the social media in the recent years in the Arab world, however, has fueled interest in sentiment analysis [31]. In this regard, there has been a relative increase in sentiment analysis studies over the recent years. Boudad, et al. [32] assert that sentiment Analysis in Arabic has grown drastically in the last five or six years. Despite these recent developments

and, the plentiful data resources on social media platforms and networks, and the growth of Arabic text content on the web, Arabic sentiment analysis is still facing serious challenges in terms of available tools and reliable systems.

In their evaluation of sentiment analysis in Arabic, Elghazaly, et al. [33] argue that sentiment analysis in Arabic is still very challenging for different reasons, including the fact that Arabic is not a case-sensitive language and does not use capital letters. Furthermore, Arabic is a highly inflected language and has a very flexible morphological system that make it very difficult for machine learning systems to perform appropriately [34-36]. Likewise, Nassif, et al. [37] stress that “despite the plentiful online Arabic content, the research on ANLP in general and sentiment analysis in particular is still suffering from the lack of tools and resources available as compared to English’. They attribute the problem mainly to the morphologically complex system of Arabic which has negative impacts on the functionality of different NLP systems including sentiment analysis.

These challenges have been clearly reflected on the sentiment analysis applications in marketing, product analysis, social media monitoring, brand reputation management, decision-making processes, and measuring public opinion in the Arab countries. This study seeks to bridge this gap in the literature through evaluating the sentiments of the Egyptian citizens on social media towards regulations and government policies as represented in the analysis of the sentiments, opinions, and attitudes of the Egyptians towards the new Real Estate Registration Law as expressed on their posts, comments, and replies on Facebook.

IV. METHODS AND PROCEDURES

To extract information about the users’ sentiment polarity (positive, neutral or negative), Facebook posts were used. The rationale is that Facebook is the most popular social media platform in Egypt according to recent statistics, as shown in Figure 1.

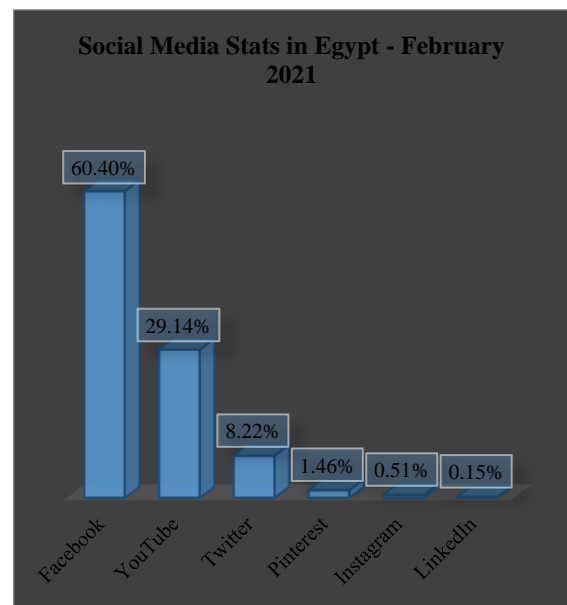


Fig. 1. Social Media Stats in Egypt - February 2021.

According to a recent report released by Facebook, Egypt had 50,220,000 Facebook users in February 2021, accounting for 47.4 percent of the country's total population. The majority of the users (62.4 percent) were male users. The largest user group was between the ages of 25 and 34. (15,800,000). Men lead by 4,200,000 users aged 25 to 34, which is the largest gap between men and women, as shown in Figure 2.

This large user base combined with a significant event such as the enforcement of a new law that has direct impacts on the life of millions of citizens makes Facebook data ideal to understand the attitudes of the Egyptian citizens towards the issue. The Arabic key terms 'الشهر العقاري' translated as 'registry office' and 'تسجيل الشهر العقاري' translated as 'real estate registration' were searched to gather relevant data over the period February 23-March 04, 2021.

The decision to use keywords rather than hashtags was based on the premise that hashtags are usually used by more experienced users who are at least somewhat familiar with the idea of hashtags and therefore have more experience than other users. We attempted to make our data sample as inclusive and representative as possible for generalizability and reliability purposes.

A corpus of 16478 documents was built. These were collected from the posts, comments on the posts, and replies to the posts on the real estate registration law. Picture posts were also transcribed and included. All texts were then tokenized and normalized for addressing variation in text length. Stemming, which is a normal procedure in standard classification systems, was not executed to make use of the linguistic richness of the Arabic derivational and inflectional morphemes [38]. A matrix was then generated including all the 16478 documents. The matrix included 60,968 words/variables.

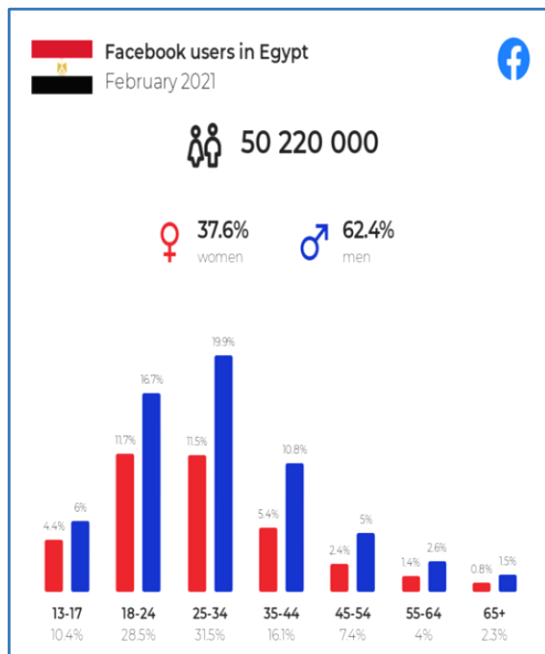


Fig. 2. Distribution of Facebook Users in Egypt as of February 2021.

One major problem with the corpus, however, was the high dimensionality of the data. This is referred to in text classification applications as the curse of dimensionality [39-42]. To address the problem, two steps were taken. First, a stop-word list with all the function words including articles, prepositions, and pronouns, was executed. Second, term weighting was carried out using Term Frequency Inverse Document Frequency (TF-IDF) to keep only the most distinctive variables within the data matrix. This is shown as follows.

In document classification applications, function words are classified irrelevant and noisy terms since they do not carry lexical significance. Accordingly, content/lexical word indexing is always recommended as an effective approach for a reliable text classification performance [43]. The premise is that the inclusion of such irrelevant variables is useless and has adverse impacts on the accuracy and reliability of classification applications [44]. It is a standard practice in text classification applications to remove function words using stop-word lists so that only bearing content words are only retained [45]. The default method for removing function words is a linguistic one. This is usually executed through the implementation of a stop-word list where all function words are identified and removed. In our case, the removal of function words through the execution of stop-word lists had the effect of reducing the matrix into 38,793 variables.

A TF-IDF analysis was then carried out to identify the most distinctive variables within the data collection, as shown in Figure 3. Based on the TF-IDF analysis, only the highest 200 variables with TF-IDF values were retained.

Text classification was finally used for classifying the selected data into three main classes/values: Positive, Negative, and Neutral. Text classification can be simply defined as the task of automatically sorting a set of documents into a number of classes or categories where each is given a label [46, 47]. Classification relies on priori reference structures that divide the space of all possible data points into a set of classes that are usually, but not necessarily, nonoverlapping [48]. A text classification task starts by discovering and finding groups that have similar content then organizing our perceptions of these groups into categories. In other words, clustering places documents into natural classes while classification places them into predefined known ones [49].

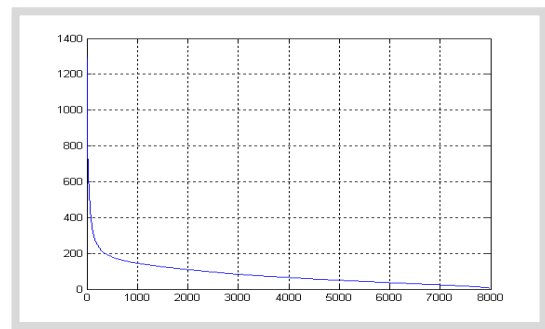


Fig. 3. Term Weighting Analysis using TF-IDF.

The classification of the data was carried out using the lexical-based model developed by Tabodada et al. in 2011. In this model, sentiment-bearing words (including adjectives, verbs, nouns, and adverbs), are extracted and used to calculate the semantic orientation of each document/post [50]. In our case, the classification of the data into the three classes: positive, negative, or neutral sentiment is achieved through a three-level analysis, as shown in Figure 4.

At the document level, also known as message level, the classifier or classification system categorizes the polarity of the post as a collective body. For example, given a product review, the system decides whether the text message is overall positive, negative, or neutral. It is assumed that the message/document conveys only one point of view on a single topic. At the sentence level, the classifier or classification system determines the polarity of each sentence contained in a post. The assumption is that each sentence, in a given post, denotes a single opinion on a single entity. Finally, entity and aspect level analysis is more comprehensive than message and sentence level analysis. It is based on the concept that an opinion is made up of two parts: a sentiment and a target of opinion [51].

As a final step, polarized words were counted. It was obvious that the number of negative word appearances is much greater than the number of positive word appearances, as shown in Figure 5.

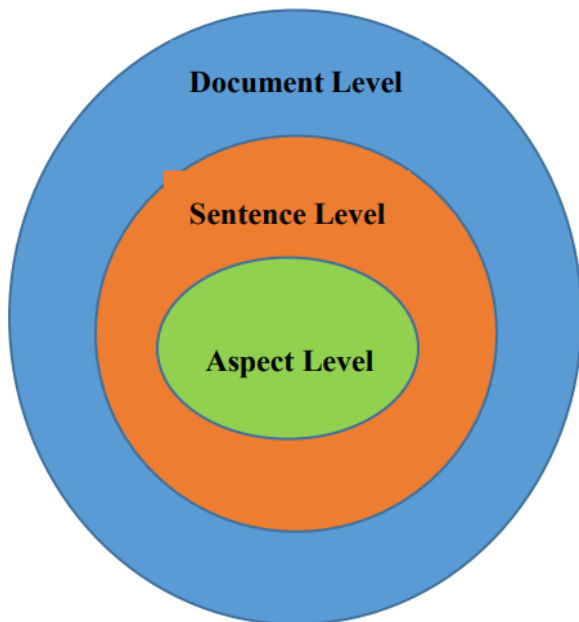


Fig. 4. Levels of Sentiment Analysis.

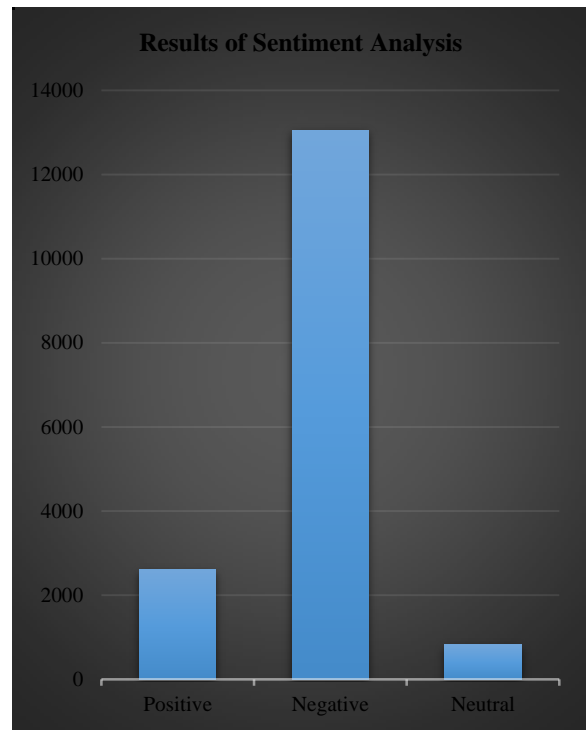


Fig. 5. Results of the Sentiment Analysis.

V. RESULTS AND DISCUSSIONS

The results indicate that the vast majority of the users (around 80%) had negative sentiments towards the new law. This is clearly reflected in the negative words including No and reject they used in reference to the law. Many of the posts referred to the law as illogical. They explained that it is illogical during the world financial crisis caused by COVID-19 to impose additional fees on the citizens. Many of the negative sentiments focused on the concept that the new law imposes heavy financial burdens and requires several bureaucratic steps. Furthermore, many posts also rejected any link between the new law and the access to the very basic facilities including gas, water, and electricity.

Positive sentiments towards the amended law, in turn, stressed that the new law regulates and protects individual ownership. It also addresses many of the inherent limitations within the old law. These implied different arguments such as “About 95 per cent of Egypt’s real estate wealth is unregistered, and that the government hopes that all citizens register their real estate units, with the objective of securing their properties”, “Registering real estate properties in registry offices helps in securing the citizens’ ownership rights, and also leads to abolishing the bad phenomenon of illegal construction and slum areas”, and “Registration contributes to increasing the value of properties”. These are shown in Fig. 6.



Fig. 6. Examples of Sentiments towards the New Law.

Referring to the research questions, it can be claimed that automatic sentiment analysis was usefully used in evaluating the sentiments of the Egyptians towards the new Real Estate Registration Law as expressed on their posts, comments, and replies on Facebook. Through the classification of the sentiments into Positive, Negative, and Neutral, it was obvious that the vast majority of the citizens were against the law. These findings represent good opportunities for policymakers and legislators to know more about people's sentiments, attitudes, and perceptions.

Despite the effectiveness of automatic sentiment analysis in providing a general idea about the perceptions and sentiments of citizens towards the issue, this automatic approach cannot deal effectively with all the data sorts collected for the purpose of the study. It was obvious that many of the posts implied judgments, arguments, agreements, disagreements, and even political stances as reflected on the declarations by the Minister of Justice on the issue who accused opposing movements of using the issue for their own agendas, as shown in Figure 7.

It was also obvious that many of the negative posts on the issue were characterized by humor. The comics below are based on Egyptian movies where users made some funny scripts on the issue, as shown in Figure 8.



Fig. 7. The Minister of Justice's Comments on the New Law.



Fig. 8. Irony-based Jokes of the New Law.

In this sense, it is challenging for automatic approaches of sentiment analysis to define accurately the sentiments of the users or individuals towards given issues.

Elghazaly, et al. [33] agree that it is still a challenging task to perform sentiment analysis appropriately in Arabic due to the fact that automatic sentiment analysis is largely context-free which is not appropriate for Arabic where the use of words and expressions is very subjective. They explain that the word "الاستين" (a spare tire) was used as negative reference to former Egyptian President Mohammed Morsi during the 2012 Presidential elections. The absence of context in sentiment analysis thus poses serious challenges for the performance of sentiment analysis systems. Nassif, et al. [37] agree that there is a need for more efforts to address the subjective data in Arabic sentiment analysis applications.

Given that automatic approaches to sentiment analysis do not consider context, it becomes important to incorporate more sophisticated language approaches including functional linguistics, Critical Discourse Analysis (CDA), and semiotics into sentiment analysis applications.

VI. CONCLUSION

This study adopted Sentiment analysis methods to evaluate the sentiments, opinions, and attitudes of the Egyptians towards the new Real Estate Registration Law as expressed on their posts, comments, and replies on Facebook. It was obvious that the majority of the citizens had negative sentiments and attitudes towards the new law. Despite the effectiveness of the automatic sentiment analysis tools of the sentiments of social media users towards the issue, these automatic tools cannot effectively evaluate many posts including judgments, arguments, and irony. In this regard, it is important to further understand the contextual aspects of the data. Although this study was limited to the evaluation of the users' opinions towards the new Real Estate Registration Law, the results have positive implications to the sentiment analysis applications for obtaining a general idea about the perceptions of citizens concerning specific issues.

ACKNOWLEDGMENT

We take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfilment of the current research project.

REFERENCES

- [1] R. Y. Shapiro and L. R. Jacobs, *The Oxford Handbook of American Public Opinion and the Media*. Oxford: Oxford University Press, 2013.
- [2] M. McCombs, *Setting the Agenda: The Mass Media and Public Opinion*. Cambridge, UK: Polity Press, 2013.
- [3] S. C. Woolley and P. N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press, 2018.
- [4] J. C. N. Raadschelders, *The Three Ages of Government: From the Person, to the Group, to the World*. Michigan: University of Michigan Press, 2020.
- [5] E. Saka, *Social Media and Politics in Turkey: A Journey through Citizen Journalism, Political Trolling, and Fake News*. London: Lexington Books, 2019.
- [6] B. Gunter, M. Elareshi, and K. Al-Jaber, *Social Media in the Arab World: Communication and Public Opinion in the Gulf States*. London: Bloomsbury Publishing, 2016.
- [7] M. Dipardo. (2014, January 24, 2014) Expedia Canada responds to angry feedback with new ads. Marketing. Available: <http://marketingmag.ca/brands/expedia-ca-responds-to-angry-social-media-feedback-with-new-ads-99039/>
- [8] D. Chaffey and P. Smith, *Digital Marketing Excellence: Planning, Optimizing and Integrating Online Marketing*. London; New York: Routledge, 2017.
- [9] S. L. Grau, *Marketing for Nonprofit Organizations: Insights and Innovations*. Oxford Oxford University Press, 2021.
- [10] Z. N. Canbolat and F. Pinarbasi, "Using Sentiment Analysis for Evaluating e-WOM: A Data Mining Approach for Marketing Decision Making," in *Exploring the Power of Electronic Word-of-Mouth in the Services Industry*: IGI Global, 2020, pp. 101-123.
- [11] E. Kauffmann, J. Peral, D. Gil, A. Ferrández, R. Sellers, and H. Mora, "A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making," *Industrial Marketing Management*, vol. 90, pp. 523-537, 2020.
- [12] A. Reyes-Menendez, J. R. Saura, and F. Filipe, "Marketing challenges in the# MeToo era: Gaining business insights using an exploratory sentiment analysis," *Heliyon*, vol. 6, no. 3, p. e03626, 2020.
- [13] P. Sánchez-Núñez, C. De Las Heras-Pedrosa, and J. I. Peláez, "Opinion mining and sentiment analysis in marketing communications: A science mapping analysis in Web of science (1998–2018)," *Social Sciences*, vol. 9, no. 3, p. 23, 2020.
- [14] J. Liu, Y. Zhou, X. Jiang, and W. Zhang, "Consumers' satisfaction factors mining and sentiment analysis of B2C online pharmacy reviews," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1-13, 2020.
- [15] A. Corallo et al., "Sentiment Analysis for Government: An Optimized Approach," in *Machine Learning and Data Mining in Pattern Recognition*, vol. 9166, P. P., Ed. (Lecture Notes in Computer Science, Cham: Springer, 2015.
- [16] R. B. Hubert, E. Estevez, A. G. Maguitman, and T. Janowski, "Examining government-citizen interactions on Twitter using visual and sentiment analysis," *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, pp. 1–10, 2018.
- [17] R. Arunachalam and S. Sarkar, "The New Eye of Government: Citizen Sentiment Analysis in Social Media," in *IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, Nagoya, Japan, 2013, pp. 23–28.
- [18] A. Morsy, "Egypt's government to amend Real Estate Registration Law, postpone enforcement till January," in *Ahram Online*, ed. Cairo, 2021.
- [19] O. Khalaf, *Egypt Real Estate Registration Crisis- Policies & Scenarios*. Istanbul: Egyptian Institute for Studies, 2011.
- [20] A. Hafiz. (2021, March 1, 2021) Anger over property tax may force Egyptian government to back down. The Arab Weekly. Available: <https://thearabweekly.com/anger-over-property-tax-may-force-egyptian-government-back-down>
- [21] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment Analysis in Social Networks*. London; New York: Morgan Kaufmann, 2017.
- [22] C. A. Iglesias and A. Moreno, *Sentiment Analysis for Social Media*. MDPI AG, 2020.
- [23] Y. Wang, *A Big Data Study on the Sentiment Analysis of Social Networks and Nonlinear System Modelling*. University of Sheffield, 2018.
- [24] H. Sahoo, *Sentiment Analysis in the UAE Social Networks Context: The Case of Emirates Telecommunication Corporation*. Abu Dhabi, Emirates: Abu Dhabi University, 2017.
- [25] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. New York: Cambridge University Press, 2014.
- [26] B. Agarwal, R. Nayak, N. Mittal, and S. Patnaik, *Deep Learning-Based Approaches for Sentiment Analysis*. Springer Singapore, 2020.
- [27] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "OCA: Opinion corpus for Arabic," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 10, pp. 2045-2054, 2011.
- [28] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014/12/01/ 2014.
- [29] M. Abdul-Mageed, M. Diab, and M. Korayem, "Subjectivity and sentiment analysis of Modern Standard Arabic," presented at the *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, June 19-24, 2011.
- [30] M. Elhawary and M. Elfeky, "Mining Arabic Business Reviews," *Proceedings of the 20 10 IEEE International Conference on Data Mining Workshops*, pp. 1108-1113, 2010.
- [31] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial," *arXiv preprint arXiv:1505.03105*, 2015.
- [32] N. Boudad, R. Faizi, R. Oulad Haj Thami, and R. Chiheb, "Sentiment analysis in Arabic: A review of the literature," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2479-2490, 2018/12/01/ 2018.

- [33] T. Elghazaly, A. Mahmoud, and H. A. Hefny, "Political sentiment analysis using twitter data," in *Proceedings of the International Conference on Internet of things and Cloud Computing*, 2016, pp. 1-5.
- [34] A. Omar, "An Evaluation of the Localization Quality of the Arabic Versions of Learning Management Systems," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, pp. 443-449, 2021.
- [35] A. Omar, "Ambiguity Resolution in Arabic Localization: The Case of Learning Management Systems," *Applied Linguistics Research Journal*, vol. 5, no. 1, pp. 1-6, 2021.
- [36] A. Omar and M. Aldawsari, "Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines," *International Journal of English Linguistics*, vol. 10, no. 3, pp. 219-228, 2020.
- [37] A. B. Nassif, A. Elnagar, I. Shahin, and S. Henno, "Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities," *Applied Soft Computing*, vol. 98, p. 106836, 2021/01/01/2021.
- [38] A. Omar and W. I. Hamouda, "The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 116-121, 2020.
- [39] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-supervised Learning*. MIT Press, 2010.
- [40] M. W. Berry and M. Castellanos, *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer London, 2007.
- [41] M. W. Berry, *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer New York, 2013.
- [42] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [43] T. Joachims, *Learning to Classify Text Using Support Vector Machines*. Springer US, 2002.
- [44] G. Ignatow and R. Mihalcea, *Text Mining: A Guidebook for the Social Sciences*. SAGE Publications, 2016.
- [45] J. Brownlee, *Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems*. Machine Learning Mastery, 2017.
- [46] M. W. Berry and J. Kogan, *Text Mining: Applications and Theory*. New York: Wiley, 2010.
- [47] J. Eisenstein, *Introduction to Natural Language Processing*. MIT Press, 2019.
- [48] H. Maranis and D. Babenko, *Algorithms of the Intelligent Web*. Greenwich: Manning Publications Co., 2009.
- [49] G. Singh, *Text Classification*. London: University College London, 2019.
- [50] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267-307, 2011.
- [51] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of Sentiment Analysis in Social Media Networks: An Overview," in *Sentiment Analysis in Social Networks* London; New York: Morgan Kaufmann, 2017, pp. 1-12.

Factors Affecting Mobile Learning Acceptance in Higher Education: An Empirical Study

Nahil Abdallah¹

School of Engineering and
Technology, Aldar College
University, Dubai, UAE

Odeh Abdallah²

Department of Fundamentals
Religion, An-Najah National
University, Nablus, Palestine

OM Bohra³

School of Business Administration,
Aldar College University
Dubai, UAE

Abstract—The use of mobile tools to support learning and teaching activities has become a significant part of the informal learning process. Mobile learning (M-learning) is used to considerably develop the forms of learning activities made by learners, and support the learning process. The effective application of M-learning in higher educational institutions, however, is based on the learners' adoption. It is therefore essential to define and investigate the factors affecting the desire of learners to use and adopt M-learning. Thus, this research investigates the factors affecting students' intention to adopt M-learning in institutions of higher education. To achieve the objectives of this research, a model is proposed based on the Unified Theory of Acceptance and Use of Technology (UTAUT) model and the Technology Acceptance Model (TAM). The instrument is developed using validated items from previous studies and shreds of literature. Data for this quantitative study are collected from undergraduate and postgraduate students. A Structural Equation Model (SEM) is used to analyze the data collected from 218 participants using a survey questionnaire. The findings show that students' intention to adopt M-learning is shaped by various variables consisting of personnel innovativeness, self-management, facilitating conditions, social influence, relative advantage, and effort expectancy. The research results also present several practical contributions and implications for M-learning adoption in terms of research and practice. Investigation of the required determinants may contribute to gain learners' adoption and is important to enhance the learning experience of students and help them improve their knowledge and academic achievement. The contribution of this paper lies in defining the factors influencing the acceptance and use of M-learning systems by students of higher education in Palestine. Hopefully, the results of the study are valuable for policy-makers in designing comprehensive M-learning systems.

Keywords—Mobile learning; UTAUT; structural equation modeling; tam; technology acceptance

I. INTRODUCTION

Advances in information and communication technology have hugely impacted our daily lives over the previous decades. These developments have recently been recognized as a potential for economic and social developments and competitiveness enhancements. It is also regarded as the most significant probable force in the twenty-first century to develop education. These developments have profoundly affected the techniques of learning and teaching and the governance of the instructional system [1].

There is a growing concern in the learning procedures, alongside innovative technological instruments, such as applied in M-learning [2]. M-learning is prescribed by [3] as the "Learning delivered to students on mobile devices such as Personal Digital Assistants (PDAs), smartphones and mobile phones". Zero technologies have geographically traversed such as mobile phones. Mobile wireless technology has been used in the classroom to improve the quality of learning at different levels and shift the way we live and thus, the way we learn begins to change [4]. Undeniably, cellphones are regarded as the lifeline of the next generation. The extensive accessibility and comparatively low prices of phone devices have strongly opened up new possibilities for leveraging the power and universality of mobile technologies to improve learning and expand instructional possibilities [5]. The rapid advancement of mobile devices and wireless networks within higher educational institutions makes university campuses an appropriate venue to incorporate student-centered M-learning [6]. In the same vein, mobile phones can extend, improve, support, and facilitate learning and teaching activities. As put by [7], innovation in mobile devices enables learners to access instructional emails, portals, library assistants, online data, and project teams. Besides, M-learning improves the flexibility of learning by adjusting learning to be more personalized [8] and helps to access the necessary subject materials in the class regardless of the limitation of time and place [9].

Though M-learning furnishes learners with significant potential capabilities [9], different problems still hinder the use of this technology, alongside other educational matters concerning the acceptance of mobile technology in schools; will this new technology be accepted by users (students and lecturers)? And may they be willing to adopt M-learning [6]. As stated by [9], M-learning success is essentially based on the readiness of learners to embrace a new technology differing from prior styles of learning. Also, [8] has stated that the key success variables of M-learning lean largely on the willingness of learners and their intellectual involvement in mobile activities. Even with the rapid development of M-learning technology, M-learning is still at the initial stages [10]. It has been observed that most of the work on M-learning is initiated in developed countries. However, the notion of mobile education is still a new venue and a rare practice in developing countries [6]. Importantly, there is a lack of empirical research findings on the variables driving the implementation of M-learning [8, 11]. It should be acquainted that at the initial phase of applying M-learning in higher education institutions,

students' views of this new technology need to be completely investigated and considered [12]. Thus, it is necessary to carry out research recognizing the necessary aspects of the adoption of M-learning.

In Palestine, M-learning has not been officially embraced in higher education institutions. For instance, the views of university teachers on incorporating mobile technology in their teaching are not taken into account by officials [13]. Despite the fast growth of mobile technologies as a new brand learning platform, the variables influencing M-learning adoption remain uncertain [14]. As asserted by [14], Mobile learning at a tertiary level is still in the beginning stages of implementation globally, and the pedagogy surrounding mobile learning is evolving and requires further pieces of research. A study conducted by [15] has indicated that almost all university students own mobile devices. Therefore, for the effective adoption of M-learning in educational institutions, numerous variables affecting learners' acceptance need to be resolved [16]. Against this, the purpose of this research is to explore the variables influencing the adoption of M-learning by college learners and point out whether the previous experience of using mobile devices impacts the adoption in various higher educational institutions. Further, a model of M-learning acceptance constructed on TAM and UTAUT is used as a theoretical basis. Nevertheless, prior studies suggest that the basic construction of UTAUT may not fully represent the particular impacts of M-learning possibly changing the behavioral intention of a user to use a mobile device [17]. For this reason, this research also examines some additional constructs considered as significant determinants of behavioral intention for M-learning.

This research article is structured as follows: part two presents a review of the literature concerning M-learning systems. Section three presents the research model and hypotheses. A description of the research methodology is given in section four. Section five presents data analysis and results, while section six offers discussion and implications. Section seven concludes this paper.

II. M-LEARNING

M-learning, as a rather evolving approach, has been enlightened with diverse descriptions in the literature [18]. It is described by prior research as an alteration of e-learning. Reference [19] has described M-learning as a new tool rapidly advanced to offer E-learning with the use of personal mobile devices. M-learning can be available in any place and at any time, together with traditional teaching settings such as classrooms, workplaces, in transit, and at home, etc [20,21].

M-learning has increasingly become essential as mobile technologies and wireless communications are rapidly developed and accepted by the concerned parties [22]. As reported by [23], M-learning can improve the whole learning procedures and educational experience. The evolution of M-learning not only delivers education through various settings but also allows students to learn at any time [24]. Villa et al., [25] have stated that mobile devices can improve the way learners cooperate and their behavioral intention towards learning, mostly because they are no longer limited by the constraints of time and space. M-learning also promotes

cooperative experiences and relationships with diversities and possibilities beyond the classroom. Therefore, the key issue of M-learning innovation is to propose learning opportunities at anytime and anywhere accomplished using various mobile devices [3]. Even though the decent and the indecent aspects of mobile devices are clear to all, it remains unclear what inspires learners to accept such technology in their learning and whether such usage has a long-term positive impact on the growth of education in general and on students' academic achievement in particular. This is where dedicated models and theories come in to elucidate the technological acceptance phenomenon [26].

III. RESEARCH MODEL AND HYPOTHESES

Despite the superb developments related to using information systems and applications, users frequently refuse to adopt such systems. Such resistance leads to financial losses and cases of frustration for organizations due to the low acceptance rates. Hence, low adoption is regarded as the key problem to information technology success implementation [27]. The main success factor for mobile learning understands the factors that lead users to implement their M-learning methods [2]. Over the years, different models and theories have been developed to support researchers in various contexts on the acceptance and use of technology.

The Unified Theory of Acceptance and Use of Technology (UTAUT) [28], and the Technology Acceptance Model (TAM) [29] are regarded as outstanding theoretical models and frameworks aimed at examining individuals' behavioral intentions and usage of Information Systems (IS) and Information Technology (IT). These frameworks have been broadly adopted in numerous IT settings such as education, online banking, shopping, and healthcare informatics [30, 31]. M-learning has its particular unique features, varying from other IS/IT settings [9]. The research, therefore, develops a contextualized model, specifically presented to inspect learners' acceptance of M-learning in higher education institutions in Palestine. The proposed research model (Fig 1) comprises seven variables. Since M-learning is not formally executed in Palestinian educational institutions, the dependent factor of this study model shall be behavioral intention (BI) rather than usage behavior. The independent factors adopted are as follows: effort expectancy (EE), relative advantage (RA), social influence (SI), facilitating conditions (FC), Personnel innovativeness (PI), and self-management (SM). Moreover, this research tests the moderating influence of using mobile experience (see Fig 1).

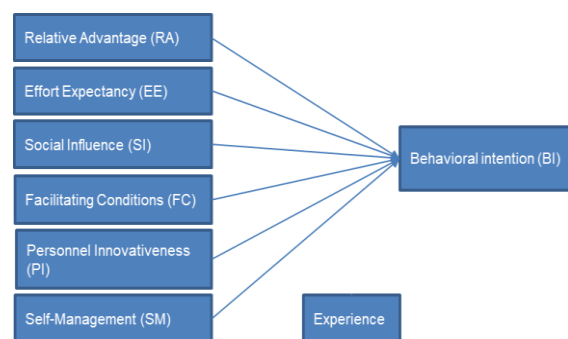


Fig. 1. Research Model.

A. Relative Advantage (RA)

Reference [32] has defined relative advantage as “the degree to which a new technology or invention is thought to be more useful than its predecessor”. RA is compared to TAM’s perceived usefulness construct, and also along the lines of UTAUT’s performance expectancy concept. RA has been frequently used by several researchers. If students perceive that it is useful to use M-learning, they will be more likely to accept it [33]. Reference [34] has mentioned that the RA of the M-learning setting emerges from the unique features of mobile phones compared to traditional learning. With characteristics like connectivity, availability, flexibility, and ubiquity, learners tend to find M-learning helpful as it enables them to use a device of their choice, and smoothly process data without any location and time constraints [9, 12, and 35].

Applying relative advantage to the M-learning setting suggests that learners might find M-learning helpful as they will learn rapidly and conveniently and enhance their learning productivity as well [36]. This empirically verified construct is a robust predictor of technology acceptance, substantially affecting the intention of end-users [4, 37, and 38]. Hence, the proposed hypothesis in this research is:

H1: “PA has a positive effect on behavioral intention to use M-learning”.

B. Effort Expectancy (EE)

Effort expectancy is described by [39] as the level of easiness and efforts required to use the technology. Ease of use and complexity (TAM2) are viewed as the two constructs from prior frameworks that are related to the conception of effort expectancy. Literature shows that M-learning adoption is largely affected by the learning system’s ease to use construct [8, 26, 40, 41]. M-learning used by learners of higher education should be simple to learn and should take a short time to comprehend [19]. According to [42], if learners regard M-learning software and hardware as user-friendly, they might be very interested in adopting it in their learning activities. Students will expect the various M-learning activities and procedures to be simple and easy to handle regardless of the limited capabilities of mobile devices. Based on UTAUT, students are anticipated to accept an M-learning depending on whether it can be used easily or not. It is therefore hypothesized that:

H2: “EE has a positive effect on behavioral intention to use M-learning”.

C. Social Influence (SI)

The social dimension investigates the impact of environmental elements such as other individual’s attitudes and social pressure executed on people. Social factors show how people, who are relevant to end-users, influence them toward accepting information technology applications [43]. Subjective norms, voluntariness, and image are grouped under a social dimension to assess the outcomes of others’ beliefs on users’ decisions to use certain technology [44].

According to [8], the uncertainty level at the early stage in technology adoption is high, where probable users tend to look for optimistic indications of convenient results from social

variables. Reference [45] has indicated that students may not be prepared to accept and embrace the new technology unless they are encouraged by other people who could influence their attitude and behavior. Many students are ready to use M-learning after recommendations from users of the technology such as their peers, workmates, friends, or lecturers [42]. Similarly, Venkatesh et al., [28] have proved that subjective norms had a durable impact on information technology acceptance decisions. Nevertheless, the influence of the subjective norm is reduced over time but is still valuable and significant [17]. Diverse studies have shown that social factor is influential in determining users’ adoption of M-learning [4, 8, 9, 17, and 46]. It is therefore hypothesized that:

H3: “SI has a positive effect on behavioral intention to use M-learning”.

D. Facilitating Conditions (FC)

As an important variable, facilitating conditions are defined by [37] as the extent to which an individual assumes an organizational and technological infrastructure exists to enable the use of a specific technology, particularly, the existing external recourses (effort, money, and time) alongside the technological resources required to enable a certain behavior.

Literature shows that suitable FCs (e.g. training, technical assistance, and adequate assets) are critical to adopt the technology [4, 38, and 47]. Environmental variables are affecting the willingness of an individual to perform the task. Several studies have found numerous complications and technical issues prohibiting students from accepting and using M-learning. Of these complications and technical issues are lack of data input capabilities, unfriendly user interfaces, limited memories, and disk capacities, lack of standardizations, low storages, low bandwidths, small screen sizes, short battery life, lower display resolutions, limited processor speeds, and less surf-ability [48]. Equally, Iqbal and Qureshi [49] have also mentioned that when learners move to M-learning, they face several technical difficulties. [50] has stated that this factor has a positive relationship with the intention to use IT/IS. In this context, the perception of the support provided by technicians’ staff and system administrators affects learner satisfaction and the decision to use the system. Therefore, facilitating conditions are found to be an important variable influencing the user’s intention and attitude. It is therefore hypothesized that:

H4: “FC has a positive effect on behavioral intention to use M-learning”.

E. Personnel Innovativeness (PI)

PI in the domain of IT is described as an individual’s tendency reflecting his or her proclivity to experiment with and adopt new IT regardless of others’ communicated experience [51]. Innovative individuals can understand the usefulness and usability of new technology applications more easily than non-innovative individuals. It is noted that people with elevated innovation have a high level of acceptance of new technology. As said by [52], an innovative person takes risks and is capable of dealing with uncertainty.

Al-Busaidi and Al-Shihi [53] have hypothesized and empirically proved that the degree of individual IT innovativeness has a substantial affirmative impact on the

attitude to accept a particular technology. Similarly, [54] has confirmed that there is a direct relationship between innovativeness and M-technology adoption. Likewise, various researchers have tested personal innovativeness predictor and found that PI has a strong effect on students' intentional behavior to accept M-learning [19, 38, and 52]. Accordingly, the present study hypothesizes that:

H5: "PI has a positive effect on behavioral intention to use M-learning".

F. Self-Management (SM)

As put by [55], self-management of learning is anticipated to be one of the central issues in the education sector because it plays an important part in promoting successful learning and acts as an indispensable driver of learning performance. Self-management is defined as the degree to which a person believes she or he is self-disciplined and participates in a highly autonomous learning environment [56].

Literature shows that self-management of learning is a significant factor in predicting M-learning acceptance [9, 44, and 57]. A student with high self-directed learning competence prefers accepting and using M-learning [48]. Students sometimes need to manage their education because they are separated from teachers, peers, colleagues, and institutions [17], which in turn requiring students to control their education [11]. Self-management is an important success factor in the development of flexible service, distance, and resource-based learning, namely: M-learning, generating a basic need from students to control their learning [58]. Hence, self-management is considered another important variable to examine university students' attitude to accept M-learning, where the following hypothesis is presented:

H6: "SM has a positive effect on behavioral intention to use M-learning".

IV. RESEARCH METHODOLOGY

Since the primary purpose of this research is concerned with the M-technology and of students' satisfaction towards M-technology learning, higher education students in Palestinian universities are considered as the unit of analysis. Notably, there is no mandatory requirement for these students to accept M-learning.

A survey instrument was developed to acquire university students' opinions. The questionnaire is divided into two sections: demographic profile of the respondents, and replies concerning the variables that is RA, EE, SI, FC, PI, and SM, and a dependent factor "Behavioral Intention" to use M-learning. In this research, the sampling method used is stratified random sampling [59]. In this sampling, the population is divided into classes called strata, and randomly selected individuals are drawn from each stratum. This implies that they should reflect the population's heterogeneity while remaining homogeneous among themselves.

There have been numerous proposed rules of thumb for the minimum sample size of structural equation models. The generally recognized representative sample parameter ratio is $N:p = 5:1$ [60]. To obtain a reliable estimate, a five-to-one response ratio is expected for each parameter. The appropriate

sample size needed to test the model's reliability is 135 with a total of 27 elements. Of all 300 questionnaires distributed to learners, 228 questionnaires have been returned. Of the returned questionnaires, 10 are described as unfinished and hence are excluded. In the end, 218 questionnaires are considered valid for further analysis, giving a response rate of 72.6%. Table 1 shows the characteristics of the respondents. A 5-point Likert scale is used to represent the responses of the subject. A 5-point Likert scale is preferred to enable respondents to answer the questions and understand better what option he/she should select for improving answers' quality.

V. DATA ANALYSIS AND RESULTS

For the assessment of the hypothesized causal relations in the proposed research model, choosing the right statistical method is crucial. Structural Equation Modelling (SEM) is a statistical modeling method used to test theoretical or abstract models. Based on reference [61], SEM allows researchers to investigate the interrelationships between multiple variables at the same time. Furthermore, it is a powerful tool that provides sophisticated statistical measures for dealing with complex frameworks.

A. Demographics and Descriptive Statistics

All the retrieved questionnaires are appropriately entered into the SPSS version 17.0 to conduct the statistical analysis. The respondent demographic profile presented in Table 1 shows that 39% of the respondents are female and almost 61% are male. Respondents aged <20 years are the largest age group, representing 51.3% of the sample size. Respondents with more than four years of experience have formed 81.1% of the sample. To locate outliers and missing values, the data are tested. All the out-coded variables are corrected and the normality of the data is also suitably checked through skewness and kurtosis. Cronbach's Alpha is also used to check data reliability for each construct. Cronbach's alpha ranges from 0 (completely unreliable) to 1 (perfectly reliable). Reference [62] has stated that the closer Cronbach's Alpha to 1.00, the higher the reliability of the measure is. The reliability result is as follows: RA (0.91), EE (0.82), SI (0.90), FC (0.79), PI (0.83), SM (0.84), and BI (0.91). Most metrics have Cronbach's alpha values greater than 0.80, indicating that they are highly reliable. As a result, there is no need to alter or modify the survey questions to increase the alpha coefficients.

B. Measurement Model

This research employed a two-phased approach to SEM analysis. First, the measurement model is estimated using CFA to assess the model's validity and reliability. Second, the structural model is used to test hypotheses between constructs. Hypothesized relationships between latent constructs are tested through the assessment of the structural model. To check if the hypothesized structural model and each of the proposed hypotheses have fitted the data, the Goodness-Of-Fit (GOF) indices, alongside the parameter estimates coefficients are examined. The reliability processes are carried out by assessing the reliability of the individual items and the composite reliability of constructs. The significance of individual item loadings is used to determine individual item reliability. The loading of each item on its underlying construct should be \geq

0.707, whereas the composite reliability (CR) should be ≥ 0.7 [63]. The loadings of each of the items on their theoretical constructs are ≥ 0.707 as shown in Table 2. Furthermore, CR values are all ≥ 0.7 .

On the other hand, the magnitude and importance of the direction between latent variables and their indicators are used to assess validity. Discriminant validity and convergent validity are used to determine the validity of the construct. According to [63], the perfect convergent validity findings are achieved when standardized loading estimates are 0.7 or higher, AVE estimation is greater than 0.5, and reliability estimation is greater than 0.7. Having followed the abovementioned suggestions, this study mainly used $0.7 > 0.5 > 0.7$ as the minimum cut-off criteria for factor loadings, AVE, and composite reliability in evaluating the convergent validity.

To measure the discriminant validity, AVE for each variable is compared with the corresponding squared inter-construct correlation (SIC). If AVE estimations are found to be consistently larger than SIC estimation, it indicates support for the discriminant validity of the construct. Table 3 shows all preceding conditions achieved by the variables.

TABLE I. DEMOGRAPHIC STATISTICS OF THE RESPONDENTS

Measure	Item	Frequency	Percentage (%)
Gender	Male	134	61.4%
	Female	84	38.6%
Age	<20	112	51.3%
	20-24	75	34.4%
	>25	41	18.8%
Experience of S Phone	< 3 Year	41	18.9%
	>4 Years	177	81.1%

TABLE II. THE MEASUREMENT MODEL ANALYSIS

Construct	Item	Loading	CR	AVE	ASV
RA	RA1	0.86	0.92	0.63	0.063
	RA2	0.88			
	RA3	0.90			
	RA4	0.83			
EE	EE1	0.77	0.89	0.63	0.031
	EE2	0.82			
	EE3	0.79			
	EE4	0.82			
SI	SI1	0.82	0.93	0.59	0.098
	SI2	0.90			
	SI3	0.84			
	SI4	0.85			
FC	FC1	0.87	0.90	0.67	0.017
	FC2	0.81			
	FC3	0.92			
	FC4	0.80			
PI	PI1	0.90	0.94	0.51	0.026
	PI2	0.91			
	PI3	0.83			
SM	SM1	0.88	0.93	0.68	0.012
	SM2	0.82			
	SM3	0.83			
	SM4	0.87			
BI	BI1	0.92	0.94	0.71	0.013
	BI2	0.91			
	BI3	0.94			

TABLE III. DISCRIMINANT VALIDITY ANALYSIS

	RA	EE	SI	FC	PI	SM	BI
RA	0.88						
EE	0.06	0.72					
SI	0.03	0.08	0.92				
FC	0.02	0.02	0.01	0.81			
PI	0.02	0.02	0.09	0.17	0.87		
SM	0.06	0.08	0.11	0.1	0.07	0.89	
BI	0.04	0.13	0.04	0.11	0.41	0.09	0.77

C. Structural Model

Table 4 indicates the structural model fit indices defining how well the presented model fits the collected data. The proposed structural model in this research is found to be valid.

Coefficient parameter estimates are another significant component of the structural model assessment. The path significance of each relationship is analyzed to test research hypotheses, and the estimated population covariance matrix for the structural model is calculated using parameter estimates. Critical ratios, standardized estimates, and the p-value are used correctly to examine the hypotheses of this study.

When the critical ratio (CR or t-value) is greater than 1.96, it is presumed that the correlation is statistically significant at the 0.05 level [63]. Based on the path estimates and CR, all of the casual paths in the model are examined. The path analysis for all the variables “RA ($\beta=0.17$), EE ($\beta=0.20$), SI ($\beta=0.23$), FC ($\beta=0.15$), PI ($\beta=0.13$), and SM ($\beta=0.14$)” has significant positive effects on BI, and therefore they are operated as key variables assisting the acceptance of M-learning. Thus, all hypotheses are supported.

D. Influences of Moderator Variable

To identify the moderating effect of the mobile experience variable, the study sample is divided into two groups: less than three years of experience, and more than three years of experience in using smartphones. Having established a sufficient model fit for both groups, multi-group analysis is employed. The t-test approach [64] is considered to identify the significant differences among path coefficients (Table 5). The results also show that the structural weights for group 1 (3 years or less) are statistically significant for all ($P < 0.05$). The structural loading values are 0.25, 0.40, 0.27, 0.22, 0.26, and 0.25, respectively. Similarly, the structural weights for group 2 (more than 3 years) are also statistically significant for all ($P < 0.05$). The structural loading values are 0.33, 0.34, 0.26, 0.35, 0.23, and 0.25, respectively.

TABLE IV. STRUCTURAL MODEL FIT INDICES

Model Fit Indices	$\chi^2/d.f$	GFI	AGFI	NFI	CFI	TLI	RMSEA
Recommended value	≤ 3.0	≥ 0.9	≥ 0.8	≥ 0.9	≥ 0.9	≥ 0.9	≤ 0.08
obtained	1.11	0.92	0.862	0.901	0.998	0.991	0.036

Hair et al., [63].

TABLE V. MODERATING EFFECTS

	3 years or less n = 41			More than 3 years n = 177		
	Estimate	t-value	P	Estimate	t-value	P
RA → BI	0.25	2.80	0.01	0.33	2.83	0.01
EE → BI	0.40	4.37	0.00	0.34	2.79	0.01
SI → BI	0.27	3.08	0.00	0.26	1.83	0.05
FC → BI	0.22	2.37	0.02	0.35	2.83	0.01
PI → BI	0.26	2.83	0.01	0.23	1.33	0.05
SM → BI	0.25	2.33	0.01	0.25	1.93	0.01

VI. DISCUSSION AND IMPLICATIONS

The findings show that the proposed model sufficiently predicts the students' behavioral intention to adopt M-learning. The results also demonstrate that relative advantage has a substantial optimistic impact on M-learning. The empirical results support the argument that the relative advantage of M-learning has a positive influence on the students' intention to adopt and accept M-learning. In a related sense of M-learning acceptance, empirical studies [6, 9, and 17] have found that relative advantage has a major effect on mobile use. These empirical results indicate that students are compelled to embrace and follow M-learning because of pre-existing beliefs based on a perceived relative advantage after considering its utility. As a result, if the utility of M-learning is recognized by potential users, it is more possible to be adopted on a large scale by students in different colleges with various academic majors.

In agreement with [7] and [48], effort complexity is considered a crucial enabler of M-learning adoption. The hypothesized relationship between EE and PI tested through hypothesis H2 (i.e. EE → PI) is found to be significant. Therefore, based on the parameter estimate results ($\beta = 0.20$, t-value = 4.37, $p = 0.001$), the proposed research hypothesis is supported. A researcher like [52] has also argued that EE primarily influences the students' usage intention. In line with the findings of previous research, statistical analysis of this study reveals that EE is a strong predictor of PI and an increase in students' perception that the easiness of M-learning would further enhance its capability toward the enhancement of education. The easier M-learning is perceived by students, the more likely it is to be used. Nevertheless, M-learning designers shall take into consideration the need for spontaneous and user-friendly interfaces.

People are more likely to engage in a certain action when they have a good outlook toward it and feel that important people think they should. M-learning behavioral intention to use is thought to be influenced by subjective norms. However, the results of parameter estimates ($\beta = 0.23$, t-value = 3.08) indicate a significant relationship between SI and PI. Therefore, this hypothesis is supported. These findings suggest that SI is a direct fundamental determinant of M-learning

acceptance. Many previous studies have shown empirical pieces of evidence of the direct impact of SI over PI in a similar domain and supported the studies' findings [4, 19, 38, 44, and 49]. Based on this conclusion, teachers should encourage and assist learners to achieve the advantages of M-learning.

Similarly, hypothesis H4 (FC → AU) suggesting "FC has a positive effect on behavioral intention to use M-learning" shows a significant result. Parameter estimate results ($\beta = 0.15$, t-value = 2.37) indicate that this hypothesis is found to be statistically significant at $p = 0.001$ level. Therefore, these results demonstrate that the students' belief to adopt M-learning is directly influenced by the availability of facilitating conditions. Compatible with the results of this study, many researchers [9, 17, 44, 48, and 49] also show a direct significant relationship between FC and PI. Therefore, M-learning suppliers should provide technical assistance and training for learners to encourage their interaction with M-learning applications. M-learning suppliers must also make sure free and sufficient wireless networks are available in universities.

Supported by [38], [52], [26], and [6], the findings also propose that the variable of Personnel innovativeness is a substantial enabler of M-learning acceptance. A student with strong personal innovativeness is ready to take risks and try the innovation. Meaning that, in the early stages of M-learning, an efficient approach to motivate learners with high innovations shall be considered, as it has a beneficial effect on expected achievements and performance expectancy [6].

Finally, Consistent with [7, 9, 11, and 48], the findings of this study show that students' intention to use M-learning is significantly influenced by their ability to control their learning. Such a finding suggests that learners with highly independent learning skills are more interested in using M-learning than learners with low self-learning skills. Besides, instructors should conscientiously deliver learning materials to support students' habit of constant self-learning and lifelong learning.

With regards to the moderating students' experience variable in using mobile, the findings show that there are important differences with regards to the impacts of this variable on students' behavioral intention to adopt the technology. Learners' experience of smartphone technology moderates the effects of the various variables on behavioral intention. Of all these factors, the social dimension factor was the strongest element of user intention. Therefore, friends and colleagues play a major role in encouraging other learners to adopt and use M-learning. Exactly, early users of M-learning could be used as an efficient means to persuade other learners to accept M-learning.

By combining the research results and mobile information systems from the perception of education literature, this study shows learners' attitudes and ability to use mobile learning in higher education in a systematic way. Although there are several pieces of research exploring the M-learning adoption in many countries, the author argues that there is a current lack of published studies examining the UTAUT2 model in Palestine.

The full potential of M-learning is unlikely to be realized without significant and first-degree M-learning adoption. In this setting, the findings of this study have pushed the boundaries of knowledge in the field of M-learning adoption by making an important impact on the literature review. This research also made a significant contribution to the theoretical concept of M-learning by understanding the factors affecting behavioral intentions to use M-learning and developing a simplified conceptual model that is used as a frame of reference by researchers, policymakers, and other higher educational institutions. At last, to successfully provide M-learning, it is always necessary to understand the factors boosting and preventing learning of such technology. The findings of this empirical research study add to the current literature on technology acceptance and introduce a novel framework for understanding and explaining key factors influencing students' acceptance of M-learning. The study model likewise launches a new foundational framework that can be evaluated by concerned administrators and educators to assess success variables for adopting M-learning.

VII. CONCLUSION

In a nutshell, understanding users' views are significant in the process of presenting new technologies. Based on the UTAUT and TAM models, this research suggests a model exploring the variables that influence the intention of university students to adopt M-learning in developing countries such as Palestine. The research model provides the means to understand what variables control students' behavioral intention to use M-learning and how that could influence future uses. The results of this research make an indispensable contribution to M-learning literature with empirical findings from a least developed country. Additionally, the result of this empirical research is hoped to help policymakers and administrators who are willing to adopt M-learning in similar contexts.

Like any other studies, this research has some constraints and limitations. Only students from two colleges have participated in this research. More public and private colleges should be included in future research to expand the sample population. Furthermore, since this study used a survey questionnaire as a research tool, potential studies using a blended approach should be considered to provide a more comprehensive understanding of M-learning implementation. Lastly, university instructors are the main backbone for accepting any new technology; their role has been largely neglected as main players for adopting M-learning. Future studies are required to test instructors' views of M-learning and demonstrate which difficulties they may expose when adopting M-learning in the teaching process.

REFERENCES

- [1] S. Papadakis and M. Kalogiannakis, "A Research Synthesis of the Real Value of Self-Proclaimed Mobile Educational Applications for Young Children," in *Mobile Learning Applications in Early Childhood Education*, ed: IGI Global, 2020, pp. 1-19.
- [2] L. Díez-Echavarría, A. Valencia, and L. Cadavid, "Mobile learning on higher educational institutions: how to encourage it?. Simulation approach," *Dyna*, vol. 85, pp. 325-333, 2018.
- [3] M. Sarrab, "M-learning in education: Omani Undergraduate students perspective," *Procedia-Social and Behavioral Sciences*, vol. 176, pp. 834-839, 2015.
- [4] M. S. Ahmed and A. Kabir, "The Acceptance of Smartphone as a Mobile Learning Tool: Students of Business Studies in Bangladesh," *Malaysian Online Journal of Educational Technology*, vol. 6, pp. 38-47, 2018.
- [5] S. Iqbal and Z. A. Bhatti, "An investigation of university student readiness towards M-learning using technology acceptance model," *The International Review of Research in Open and Distributed Learning*, vol. 16, 2015.
- [6] A. Abu-Al-Aish and S. Love, "Factors influencing students' acceptance of M-learning: An investigation in higher education," *The International Review of Research in Open and Distributed Learning*, vol. 14, 2013.
- [7] Y. F. Chye, J. C. Ong, J. X. Tan, and S. J. Thum, "The fundamental factors that influencing mobile learning acceptance in higher education institution," *UTAR*, 2014.
- [8] M. Sarrab, I. Al Shibli, and N. Badursha, "An empirical study of factors driving the adoption of mobile learning in Omani higher education," *The International Review of Research in Open and Distributed Learning*, vol. 17, 2016.
- [9] A. S. Al-Adwan, A. Al-Madadha, and Z. Zvirzdinaite, "Modeling students' readiness to adopt mobile learning in higher education: An empirical study," *International Review of Research in Open and Distributed Learning*, vol. 19, 2018.
- [10] A. Lu, Q. Chen, Y. Zhang, and T. Chang, "Investigating the Determinants of Mobile Learning Acceptance in Higher Education Based on UTAUT," in *2016 International Computer Symposium (ICS)*, 2016, pp. 651-655.
- [11] Al-Adwan, A. Samed, Al-Adwan, Berger, and Hilary, "Solving the mystery of mobile learning adoption in higher education," *International Journal of Mobile Communications*, vol. 16, pp. 24-49, 2018.
- [12] J. R. Batmetan and V. R. Palilingan, "Higher education students' behaviour to adopt mobile learning," in *IOP Conference Series: Materials Science and Engineering*, 2018, p. 012067.
- [13] K. Shraim and H. Crompton, "Perceptions of Using Smart Mobile Devices in Higher Education Teaching: A Case Study from Palestine," *Contemporary Educational Technology*, vol. 6, pp. 301-318, 2015.
- [14] S. I. Senaratne and S. M. Samarasinghe, "Factors Affecting the Intention to Adopt M-learning," *International Business Research*, vol. 12, pp. 150-164, 2019.
- [15] A. Z. Shaqour, "Students' Readiness towards M-learning: A Case Study of Pre-Service Teachers in Palestine," *Journal of Educational and Social Research*, vol. 4, p. 19, 2014.
- [16] T. Thomas, L. Singh, and K. Gaffar, "The utility of the UTAUT model in explaining mobile learning adoption in higher education in Guyana," *International Journal of Education and Development using ICT*, vol. 9, 2013.
- [17] N. G. Uğur, T. Koç, and M. Koç, "An analysis of mobile learning acceptance by college students," *Journal of educational and instructional studies in the world*, vol. 6, pp. 1-11, 2016.
- [18] H. Akour, "Determinants of mobile learning acceptance: an empirical investigation in higher education," *Oklahoma State University*, 2009.
- [19] Y. S. Poong, S. Yamaguchi, and J.-i. Takada, "Investigating the drivers of mobile learning acceptance among young adults in the World Heritage town of Luang Prabang, Laos," *Information Development*, vol. 33, pp. 57-71, 2017.
- [20] R. Petley, J. Attewell, and C. Savill-Smith, "Not just playing around: the MoLeNET experience of using games technologies to support teaching and learning," in *Wireless Technologies: Concepts, Methodologies, Tools and Applications*, ed: IGI Global, 2012, pp. 1429-1442.
- [21] M. Kalogiannakis and S. Papadakis, "Evaluating pre-service kindergarten teachers' intention to adopt and use tablets into teaching practice for natural sciences," *International Journal of Mobile Learning and Organisation*, vol. 13, pp. 113-127, 2019.
- [22] M. A. Embi and N. M. Nordin, "Mobile learning: Malaysian initiatives and research findings," *Malaysia: Centre for Academic Advancement, Universiti Kebangsaan Malaysia*, pp. 1-131, 2013.

- [23] B. Oberer and A. Erkollar, "Mobile learning in higher education: a marketing course design project in Austria," *Procedia-Social and Behavioral Sciences*, vol. 93, pp. 2125-2129, 2013.
- [24] G. W.-H. Tan, K.-B. Ooi, L.-Y. Leong, and B. Lin, "Predicting the drivers of behavioral intention to use mobile learning: A hybrid SEM-Neural Networks approach," *Computers in Human Behavior*, vol. 36, pp. 198-213, 2014.
- [25] E. Villa, L. Ruiz, A. Valencia, and E. Picón, "Electronic commerce: factors involved in its adoption from a bibliometric analysis," *Journal of theoretical and applied electronic commerce research*, vol. 13, pp. 39-70, 2018.
- [26] R. Vrana, "Acceptance of mobile technologies and M-learning in higher education learning: an explorative study at the Faculty of Humanities and Social Science at the University of Zagreb," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0738-0743.
- [27] M. S. Abbasi, A. Tarhini, M. Hassouna, and F. Shah, "SOCIAL, ORGANIZATIONAL, DEMOGRAPHY AND INDIVIDUALS' TECHNOLOGY ACCEPTANCE BEHAVIOUR: A CONCEPTUAL MODEL," *European Scientific Journal*, ESJ, vol. 11, 2015.
- [28] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425-478, 2003.
- [29] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319-340, 1989.
- [30] H. Celik, "Customer online shopping anxiety within the Unified Theory of Acceptance and Use Technology (UTAUT) framework," *Asia Pacific Journal of Marketing and Logistics*, vol. 28, pp. 278-307, 2016.
- [31] M. AlKailani, "Factors Affecting the Adoption of Internet Banking in Jordan: An Extended TAM Model," *Journal of Marketing Development & Competitiveness*, vol. 10, 2016.
- [32] E. Roggers, "Complex adaptive systems and the diffusion of innovation," *The Innovation Journal: The Public Sector Innovation Journal*, vol. 10, 2005.
- [33] G. Jackman, "Investigating the factors influencing students' accepting mobile learning: the cave hill campus experience," *Caribbean Educational Research Journal*, vol. 2, p. 14.32, 2014.
- [34] I. Arpaci, "A comparative study of the effects of cultural differences on the adoption of mobile learning," *British Journal of Educational Technology*, vol. 46, pp. 699-712, 2015.
- [35] B. Klimova, "Impact of Mobile Learning on Students' Achievement Results," *Education Sciences*, vol. 9, p. 90, 2019.
- [36] Y. S. Wang, M. C. Wu, and H. Y. Wang, "Investigating the determinants and age and gender differences in the acceptance of mobile learning," *British journal of educational technology*, vol. 40, pp. 92-118, 2009.
- [37] V. Venkatesh, J. Y. Thong, and X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS quarterly*, vol. 36, pp. 157-178, 2012.
- [38] A. A. Arain, Z. HUSSAIN, M. VIGHIO, and W. RIZVI, "Factors Influencing Acceptance of Mobile Learning by Higher Education Students in Pakistan," *Sindh University Research Journal-SURJ (Science Series)*, vol. 50, pp. 141-146, 2018.
- [39] V. Venkatesh, J. Y. Thong, and X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology," *MIS quarterly*, pp. 157-178, 2012.
- [40] F. Ozdamli and H. Uzunboylu, "M-learning adequacy and perceptions of students and teachers in secondary schools," *British Journal of Educational Technology*, vol. 46, pp. 159-172, 2015.
- [41] M. Alnabhan and Y. Aljaraidh, "Collaborative M-learning Adoption Model: A Case Study for Jordan," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 9, pp. 4-10, 2014.
- [42] J. Mtebe and R. Raisamo, "Investigating students' behavioural intention to adopt and use mobile learning in higher education in East Africa," *International Journal of Education and Development using ICT*, vol. 10, 2014.
- [43] L. Rashotte, "Social influence," *The Blackwell encyclopedia of sociology*, 2007.
- [44] L. Aofan, C. Qianqian, Y. Zhang, and T. Chang, "Investigating the Determinants of Mobile Learning Acceptance in Higher Education Based on UTAUT," in *International Computer Symposium*, Chiayi, Taiwan, 2016.
- [45] A. A. Taiwo and A. G. Downe, "The theory of user acceptance and use of technology (UTAUT): A meta-analytic review of empirical findings," *Journal of Theoretical & Applied Information Technology*, vol. 49, 2013.
- [46] M. Rehman, M. Anjum, F. Askri, M. Kamran, and V. Esichaikul, "Mobile learning adoption framework: An empirical investigation from learners perspective," *Journal of Quality and Technology Management*, vol. 12, pp. 1-43, 2016.
- [47] A. Aypay, H. C. Celik, A. Aypay, and M. Sever, "Technology Acceptance in Education: A Study of Pre-Service Teachers in Turkey," *Turkish Online Journal of Educational Technology-TOJET*, vol. 11, pp. 264-272, 2012.
- [48] Y. Huan, X. Li, M. Aydeniz, and T. Wyatt, "Mobile learning adoption: An empirical investigation for engineering education," *International Journal of Engineering Education*, vol. 31, pp. 1081-1091, 2015.
- [49] S. Iqbal and I. A. Qureshi, "M-learning adoption: A perspective from a developing country," *The International Review of Research in Open and Distributed Learning*, vol. 13, pp. 147-164, 2012.
- [50] M. Abbad, "Proposed model of e-learning acceptance," in *International Conference on Education and e-Learning Innovations*, 2012, pp. 1-9.
- [51] N. Schillewaert, M. J. Ahearne, R. T. Frambach, and R. K. Moenaert, "The adoption of information technology in the sales force," *Industrial Marketing Management*, vol. 34, pp. 323-336, 2005.
- [52] A. Bombaas, "Student's Intentions to Use M-learning: An Empirical Perspective from the Philippines," *Business and Economic Research*, vol. 8, pp. 68-83, 2018.
- [53] K. A. Al-Busaidi and H. Al-Shihi, "Instructors' acceptance of learning management systems: A theoretical framework," *Communications of the IBIMA*, vol. 2010, pp. 1-10, 2010.
- [54] C.-H. Liu and Y.-M. Huang, "An empirical investigation of computer simulation technology acceptance to explore the factors that affect user intention," *Universal Access in the Information Society*, vol. 14, pp. 449-457, 2015.
- [55] R.-T. Huang, "Exploring the moderating role of self-management of learning in mobile English learning," 2014.
- [56] P. J. Smith, K. L. Murphy, and S. E. Mahoney, "Towards identifying factors underlying readiness for online learning: An exploratory study," *Distance education*, vol. 24, pp. 57-67, 2003.
- [57] J. N. Lowenthal, "Using mobile learning: Determinates impacting behavioral intention," *The Amer. Jnl. of Distance Education*, vol. 24, pp. 195-206, 2010.
- [58] S. Yang, "Understanding undergraduate students' adoption of mobile learning model: A perspective of the extended UTAUT2," *Journal of convergence information technology*, vol. 8, p. 969, 2013.
- [59] G. Kalton and K. Graham, *Introduction to survey sampling* vol. 35: Sage, 1983.
- [60] R. B. Kline, *Principles and practice of structural equation modeling*: Guilford publications, 2015.
- [61] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, "Multivariate data analysis (Vol. 6): Pearson Prentice Hall Upper Saddle River," ed: NJ, 2006.
- [62] U. Sekaran and R. Bougie, *Research methods for business: A skill building approach*: John Wiley & Sons, 2016.
- [63] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, and R. L. Tatham, *Multivariate data analysis* vol. 6: Pearson Prentice Hall Upper Saddle River, NJ, 2006.
- [64] J. F. Hair, M. Sarstedt, C. M. Ringle, and J. A. Mena, "An assessment of the use of partial least squares structural equation modeling in marketing research," *Journal of the academy of marketing science*, vol. 40, pp. 414-433, 2012.

Novel Properties for Total Strong - Weak Domination Over Bipolar Intuitionistic Fuzzy Graphs

As'ad Mahmoud As'ad Alnaser
Department of Applied Science
Ajloun College, Al-Balqa Applied University
Jordan

Abstract—Through this research study, we introduced and discussed total strong (weak) domination concept of bipolar intuitionistic fuzzy graphs and in define strong domination bipolar intuitionistic fuzzy graph also strong domination. Theorems, examples and some properties of these concept are discussed.

Keywords—Fuzzy sets; bipolar intuitionistic fuzzy sets; strong (weak) bipolar intuitionistic fuzzy sets; total strong (weak) bipolar intuitionistic number

I. INTRODUCTION

The theory of fuzzy sets was introduced by Zadeh [1]. On this notion the researchers emphasized their applications in different areas such as electrical engineering economics, Computer Science, social networks, system analysis and mathematics, many researchers using this concept to generalized and study some topics [2-8]. Atanassov [9] generalized the idea of fuzzy set and gave new Concept which intuitionistic fuzzy sets. Many researchers have benefited from this new Concept in developing many old Concepts in many fields of Science [10-13]. Zhang [14] initiated a bipolar fuzzy set concept as a development to the fuzzy set theory, since a set of bipolar fuzzy is an extension of fuzzy set of Zadeh's whose membership degree range is $[-1,1]$. Also, many researchers have used this notion to study many properties [15-18]. E zhilmaran and sankar [19 - 20] have introduced bipolar intuitionistic fuzzy set and studied it on graph theory, A.Alnaser et.al in 2020 [21] used this concept to graph theory also. The concept of bipolar intuitionistic fuzzy set is considered a new and important concept as it has entered many sciences such as networks and engineering, mathematics, control systems, medicine and other sciences. In our future study we will use this concept to develop some of the results reached in many research papers such as [22 – 23].

In this paper, we will introduce and discuss total strong (weak) domination concept of bipolar intuitionistic fuzzy graphs and define strong domination bipolar intuitionistic fuzzy graph & strong domination. Theorems, examples and some properties of these concepts will also be discussed.

II. PRELIMINARIES

Definition 2.1: [1] Let G be a set, a fuzzy set δ on G just function $\delta: G \rightarrow [0,1]$.

Definition 2.2: [3] A fuzzy set δ is said to be fuzzy relation on G if the map $\gamma: G \times G \rightarrow [0,1]$ satisfy $\gamma(a, d) \leq$

$\min\{\delta(a), \delta(d)\}$ for all $a, d \in G$. A fuzzy relation is symmetric if $\gamma(a, d) = \gamma(d, a)$ for all $a, d \in G$.

Definition 2.3: [14] If $G \neq \emptyset$. A bipolar fuzzy set λ of G is object with form $\varphi = \{(i, \lambda^+(i), \lambda^-(i); i \in G)\}$ such that $\lambda^+: G \rightarrow [0, 1]$ and $\lambda^-: G \rightarrow [-1, 0]$ are mappings.

Definition 2.4: [9] If G is an non empty set. An intuitionistic fuzzy set $\mathfrak{I} = \{(k; \mu(k), \lambda(k); k \in G)\}$ such that $\mu: G \rightarrow [0, 1]$ and $\lambda: G \rightarrow [0, 1]$ are mapping such that $0 \leq \mu(k) + \lambda(k) \leq 1$.

Definition 2.5: [24] An ordered pair $G^* = (V, E)$ is graph such that V the vertices set in G^* & E the edge set in G^* .

Remark 2.6: [24] 1) If c and e are two vertices in G^* then its called adjacent of G^* when (c, e) is edge of G^* .

2) An undirected graph which has at most one edge between any two different vertices no loops called simple graph.

Definition 2.7: [24] A sub graph of G^* is a graph $S = (W, F)$ such that $W \leq V$ and $F \leq E$.

Definition 2.8: [24] $(G^*)^c$ is complementary graph of a simple graph with the same vertices of G^* .

Remark 2.9: [24] Two vertices are adjacent in $(G^*)^c$ iff they are not adjacent in G^* .

III. MEAN RESULTS

Definition 3.1: [19] If $G \neq \emptyset$. A bipolar intuitionistic fuzzy sets $\mathfrak{I} = \{(e) \mu^+(e), \mu^-(e), \lambda^+(e), \lambda^-(e); e \in G\}$ such that $\mu^+: G \rightarrow [0, 1]$, $\mu^-: G \rightarrow [-1, 0]$, $\lambda^+: G \rightarrow [0, 1]$, $\lambda^-: G \rightarrow [-1, 0]$. Are mapping, where $0 \leq \mu^+(e) + \lambda^+(e) \leq 1, -1 \leq \mu^-(e) + \lambda^-(e) \leq 0$.

Using the degree of positive membership $\mu^+(i)$ to represent the degree of satisfaction of "e" to the corresponding of property of a bipolar intuitionistic fuzzy sets \mathfrak{I} also the negative degree of membership $\mu^-(e)$ for represent the satisfaction degree of "e" for any implicit counter property corresponding for a bipolar intuitionistic fuzzy sets. By the same cases, we use the degree of positive non membership $\lambda^+(e)$ for represent the satisfaction degree of "e" to the property corresponding for a bipolar intuitionistic fuzzy sets also, the degree of negative non membership $\lambda^-(e)$ for represent the satisfaction degree "e" to some implicit counter property corresponding for a bipolar intuitionistic fuzzy sets.

If $\mu^+(e) \neq 0, \mu^-(e) = 0$ and $\lambda^+(e) = 0, \lambda^-(e) = 0$ the situation that "e" regarded as have only a positive membership property in bipolar intuitionistic fuzzy sets. While if $\mu^+(e) = 0, \mu^-(e) \neq 0$ also $\lambda^+(e) = 0, \lambda^-(e) = 0$, then it's the situation that "e" regarded as have a negative membership property. While if $\mu^+(e) = 0, \mu^-(e) = 0$ also $\lambda^+(e) \neq 0, \lambda^-(e) = 0$, it's the situation that "e" regarded as have only a positive non membership property. If $\mu^+(e) = 0, \mu^-(e) = 0$ also $\lambda^+(e) = 0, \lambda^-(e) \neq 0$ it's a situation that "e" regarded as have only the negative non membership property. It is possible for an element e to be such that $\mu^+(e) \neq 0, \mu^-(e) \neq 0$ also $\lambda^+(e) \neq 0, \lambda^-(e) \neq 0$ when a membership and non membership function of a property overlaps with its counter properties over some one portion of "e".

Definition 3.2: [19] If G is a non empty set. Then the mapping $\mathfrak{F} = (\mu_{\mathfrak{F}}^+, \mu_{\mathfrak{F}}^-, \lambda_{\mathfrak{F}}^+, \lambda_{\mathfrak{F}}^-): G \times G \rightarrow ([0, 1] \times [-1, 0] \times [0, 1] \times [-1, 0])$ is a bipolar intuitionistic fuzzy relation on G , where $\mu_{\mathfrak{F}}^+(i, j) \in [0, 1], \mu_{\mathfrak{F}}^-(i, j) \in [-1, 0], \lambda_{\mathfrak{F}}^+(i, j) \in [0, 1]$ and $\lambda_{\mathfrak{F}}^-(i, j) \in [-1, 0]$.

Definition 3.3: [19] Let $\mathfrak{F}_1 = (\mu_{\mathfrak{F}_1}^+(e), \mu_{\mathfrak{F}_1}^-(e), \lambda_{\mathfrak{F}_1}^+(e), \lambda_{\mathfrak{F}_1}^-(e))$ also $\mathfrak{F}_2 = (\mu_{\mathfrak{F}_2}^+(e), \mu_{\mathfrak{F}_2}^-(e), \lambda_{\mathfrak{F}_2}^+(e), \lambda_{\mathfrak{F}_2}^-(e))$ be two bipolar intuitionistic fuzzy sets on G . \mathfrak{F}_1 is a bipolar intuitionistic fuzzy relation for \mathfrak{F}_2 if

- 1) $\mu_{\mathfrak{F}_1}^+(e, f) \leq \min\{\mu_{\mathfrak{F}_2}^+(e), \mu_{\mathfrak{F}_2}^+(f)\}$
- 2) $\mu_{\mathfrak{F}_1}^-(e, f) \geq \max\{\mu_{\mathfrak{F}_2}^-(e), \mu_{\mathfrak{F}_2}^-(f)\}$
- 3) $\lambda_{\mathfrak{F}_1}^+(e, f) \geq \max\{\lambda_{\mathfrak{F}_2}^+(e), \lambda_{\mathfrak{F}_2}^+(f)\}$
- 4) $\lambda_{\mathfrak{F}_1}^-(e, f) \leq \min\{\lambda_{\mathfrak{F}_2}^-(e), \lambda_{\mathfrak{F}_2}^-(f)\}$
- 5) $\forall e, f \in G$.

Remark 3.4: A bipolar intuitionistic fuzzy relation for \mathfrak{F}_1 on G is said to be symmetric when $\mu_{\mathfrak{F}_1}^+(e, f) = \mu_{\mathfrak{F}_1}^+(f, e), \mu_{\mathfrak{F}_1}^-(e, f) = \mu_{\mathfrak{F}_1}^-(f, e)$ also $\lambda_{\mathfrak{F}_1}^+(e, f) = \lambda_{\mathfrak{F}_1}^+(f, e), \lambda_{\mathfrak{F}_1}^-(e, f) = \lambda_{\mathfrak{F}_1}^-(f, e)$.
 $\forall e, f \in G$.

Definition 3.5: [19] A bipolar intuitionistic fuzzy graph of $G^* = (V, E)$ that is a pair $G = (X, Y)$ such that $\mathfrak{F}_1 = (\mu_{\mathfrak{F}_1}^+, \mu_{\mathfrak{F}_1}^-, \lambda_{\mathfrak{F}_1}^+, \lambda_{\mathfrak{F}_1}^-)$ is a bipolar intuitionistic fuzzy sets in V and $\mathfrak{F}_2 = (\mu_{\mathfrak{F}_2}^+, \mu_{\mathfrak{F}_2}^-, \lambda_{\mathfrak{F}_2}^+, \lambda_{\mathfrak{F}_2}^-)$ is a bipolar intuitionistic fuzzy set of $V \times V$ such that.

- 1) $\mu_{\mathfrak{F}_2}^+(e, f) \leq \min\{\mu_{\mathfrak{F}_1}^+(e), \mu_{\mathfrak{F}_1}^+(f)\}, \forall ef \in V \times V$
- 2) $\mu_{\mathfrak{F}_2}^-(e, f) \geq \max\{\mu_{\mathfrak{F}_1}^-(e), \mu_{\mathfrak{F}_1}^-(f)\}, \forall ef \in V \times V$
- 3) $\lambda_{\mathfrak{F}_2}^+(e, f) \geq \max\{\lambda_{\mathfrak{F}_1}^+(e), \lambda_{\mathfrak{F}_1}^+(f)\}, \forall ef \in V \times V$
- 4) $\lambda_{\mathfrak{F}_2}^-(e, f) \leq \min\{\lambda_{\mathfrak{F}_1}^-(e), \lambda_{\mathfrak{F}_1}^-(f)\}, \forall ef \in V \times V$
- 5) $\mu_{\mathfrak{F}_2}^+(e, f) = \mu_{\mathfrak{F}_2}^-(e, f) = 0 \forall ef \in V \times V - E$
- 6) $\lambda_{\mathfrak{F}_2}^+(e, f) = \lambda_{\mathfrak{F}_2}^-(e, f) = 0 \forall ef \in V \times V - E$

Through this article, G^* and G is representing a crisp to a graph and bipolar intuitionistic fuzzy graph respectively.

Definition 3.6: Let $F = (\mathfrak{F}_1, \mathfrak{F}_2)$ be a bipolar intuitionistic fuzzy graphs, where

$\mathfrak{F}_1 = (\mu_{\mathfrak{F}_1}^+, \mu_{\mathfrak{F}_1}^-, \lambda_{\mathfrak{F}_1}^+, \lambda_{\mathfrak{F}_1}^-)$ and $\mathfrak{F}_2 = (\mu_{\mathfrak{F}_2}^+, \mu_{\mathfrak{F}_2}^-, \lambda_{\mathfrak{F}_2}^+, \lambda_{\mathfrak{F}_2}^-)$ are two bipolar intuitionistic fuzzy set on a non empty set V and $E \leq V \times V$ respectively.

The positive degree of a vertex $\mu_{\mathfrak{F}_1}^+, \lambda_{\mathfrak{F}_1}^+ \in G$ is $D^+(\mu_{\mathfrak{F}_1}^+(e), \lambda_{\mathfrak{F}_1}^+(e)) = \sum_{ef \in E} \mu_{\mathfrak{F}_2}^+(ef) + \sum_{ef \in E} \lambda_{\mathfrak{F}_2}^+(ef)$
Similarly, the negative degree of a vertex $\mu_{\mathfrak{F}_1}^-, \lambda_{\mathfrak{F}_1}^- \in G$ is $D^-(\mu_{\mathfrak{F}_1}^-(e), \lambda_{\mathfrak{F}_1}^-(e)) = \sum_{ef \in E} \mu_{\mathfrak{F}_2}^-(ef) + \sum_{ef \in E} \lambda_{\mathfrak{F}_2}^-(ef)$.

The degree of the vertex is $D(\mu, \lambda) = (D^+(\mu, \lambda), D^-(\mu, \lambda))$.

Definition 3.7: If $G = (\mathfrak{F}_1, \mathfrak{F}_2)$ is bipolar intuitionistic fuzzy graphs. Then the bipolar intuitionistic fuzzy graph G order given by $O(G) = (\sum_{e \in V} \mu_{\mathfrak{F}_1}^+(e), \sum_{f \in V} \lambda_{\mathfrak{F}_1}^+(f), \sum_{e \in V} \mu_{\mathfrak{F}_1}^-(e), \sum_{e \in V} \lambda_{\mathfrak{F}_1}^-(f))$.

The size of a bipolar intuitionistic fuzzy graph G is $S(G) = (\sum_{ij \in V} \mu_{\mathfrak{F}_1}^+(ij), \sum_{ij \in V} \mu_{\mathfrak{F}_1}^-(ij), \sum_{ij \in V} \lambda_{\mathfrak{F}_1}^+(ij), \sum_{ij \in V} \lambda_{\mathfrak{F}_1}^-(ij))$

Definition 3.8: if $G = (\mathfrak{F}_1, \mathfrak{F}_2)$ is a bipolar intuitionistic fuzzy graph, for each node in G has the same closed degree of neighborhood, thus G said to be totally bipolar intuitionistic fuzzy graphs. The closed degree of neighborhood of a node "e" defined as $Deg(e) = (Deg^+(e) + Deg^-(e))$, such that $Deg^+(e) = (Deg \mu^+(e) + \mu_{\mathfrak{F}_1}^+(e), Deg \lambda^+(e) + \lambda_{\mathfrak{F}_1}^+(e))$

$Deg^-(e) = (Deg \mu^-(e) + \mu_{\mathfrak{F}_1}^-(e), Deg \lambda^-(e) + \lambda_{\mathfrak{F}_1}^-(e))$.

Definition 3.9: A bipolar intuitionistic fuzzy graphs $G = (\mathfrak{F}_1, \mathfrak{F}_2)$ is called strong bipolar intuitionistic fuzzy graph, when $\mu_{\mathfrak{F}_2}^+(ij) = \min\{\mu_{\mathfrak{F}_1}^+(i), \mu_{\mathfrak{F}_1}^+(j)\}$, $\lambda_{\mathfrak{F}_2}^+(ij) = \max\{\lambda_{\mathfrak{F}_1}^+(i), \lambda_{\mathfrak{F}_1}^+(j)\}$ and $\mu_{\mathfrak{F}_2}^-(ij) = \min\{\mu_{\mathfrak{F}_1}^-(i), \mu_{\mathfrak{F}_1}^-(j)\}$, $\lambda_{\mathfrak{F}_2}^-(ij) = \max\{\lambda_{\mathfrak{F}_1}^-(i), \lambda_{\mathfrak{F}_1}^-(j)\} \forall ij \in E$.

Definition 3.10: if $G = (\mathfrak{F}_1, \mathfrak{F}_2)$ is a bipolar intuitionistic fuzzy graph, if I and J are two vertices. Then I is totally strong dominates J (J totally weak dominates I)

If

- 1) $\mu_{\mathfrak{F}_2}^+(ef) = \min\{\mu_{\mathfrak{F}_1}^+(e), \mu_{\mathfrak{F}_1}^+(f)\}, \mu_{\mathfrak{F}_2}^-(ef) = \min\{\mu_{\mathfrak{F}_1}^-(e), \mu_{\mathfrak{F}_1}^-(f)\}$ and $\lambda_{\mathfrak{F}_2}^+(ef) = \max\{\lambda_{\mathfrak{F}_1}^+(e), \lambda_{\mathfrak{F}_1}^+(f)\}$ and $\lambda_{\mathfrak{F}_2}^-(ef) = \max\{\lambda_{\mathfrak{F}_1}^-(e), \lambda_{\mathfrak{F}_1}^-(f)\}, \forall ef \in E$

- 2) $D^-(I) \geq D^-(J)$ and
- 3) Every vertex in G dominates I

Definition 3.11: If G is a bipolar intuitionistic fuzzy graph; τ_B is called total strong (weak) dominating a bipolar intuitionistic set of G if

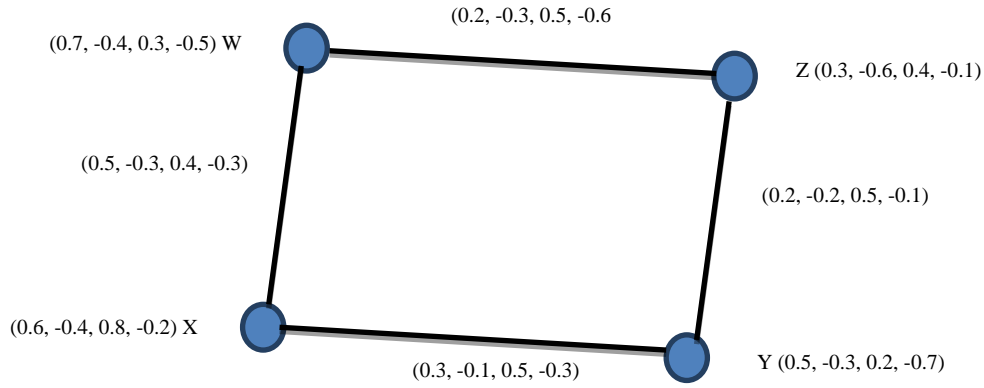


Fig. 1. Bipolar Intuitionistic Fuzzy Graph.

- 1) $\mu(H, S) \geq \mu^\infty(H, S), \lambda(H, S) \geq \lambda^\infty(H, S) \forall H, S \in V(G)$
- 2) $D^-(H) \geq D^-(S) \forall H \in \tau_B, S \in V - \tau_B$
- 3) $\mu_{\tau_B}^+(ef) = \min\{\mu_{\tau_B}^+(e), \mu_{\tau_B}^+(f)\}$
 $\mu_{\tau_B}^-(ef) = \min\{\mu_{\tau_B}^-(e), \mu_{\tau_B}^-(f)\}$ and
 $\lambda_{\tau_B}^+(ef) = \max\{\lambda_{\tau_B}^+(e), \lambda_{\tau_B}^+(f)\}$
 $\lambda_{\tau_B}^-(ef) = \max\{\lambda_{\tau_B}^-(e), \lambda_{\tau_B}^-(f)\}. \forall ef \in E$

4) τ_B is the total dominating a bipolar intuitionistic set

Definition 3.12: τ_B of a fuzzy graphs G is called minimal total strong (weak) dominating bipolar intuitionistic set of G , if there doesn't exist any total strong (weak) dominating bipolar intuitionistic set of G , whose cardinality is less than the cardinality τ_B .

Definition 3.13: A total strong (weak) dominating bipolar intuitionistic set in G is the fuzzy cardinality of minimum among all minimal total strong (weak) dominating bipolar intuitionistic set G .

Remark 3.14: The total strong (weak) domination bipolar intuitionistic number is represented by $\pi_{\tau_B}(G)$.

Example 3.15: If G is a bipolar intuitionistic fuzzy graph given by the Fig. 1.

Total strong (weak) dominating bipolar intuitionistic set $\tau_B = \{x, y\}$

Total strong (weak) bipolar domination number $\pi_{\tau_B}(\mu_{\tau_B}^+, \mu_{\tau_B}^-, \lambda_{\tau_B}^+, \lambda_{\tau_B}^-) = (1.1, -0.7, 1, -0.9)$

$Deg(\mu_x^+, \mu_x^-, \lambda_x^+, \lambda_x^-) = (0.8, -0.4, 0.9, -0.6)$

$Deg(\mu_y^+, \mu_y^-, \lambda_y^+, \lambda_y^-) = (0.5, -0.3, 1, -0.4)$

$Deg(\mu_z^+, \mu_z^-, \lambda_z^+, \lambda_z^-) = (0.4, -0.5, 1, -0.7)$

$Deg(\mu_w^+, \mu_w^-, \lambda_w^+, \lambda_w^-) = (0.7, -0.6, 0.9, -0.9)$

Order of bipolar intuitionistic fuzzy graph $O(G) = (2.1, -1.7, 1.7, -1.5)$

Size of bipolar intuitionistic fuzzy graph $S(G) = (1.2, -0.9, 1.9, -1.3)$

Theorem 3.16: If G is bipolar intuitionistic fuzzy graph also if τ_B is minimal total strong (weak) dominating. Then for each $J \in \tau_B$, if one of following axioms hold

- 1) There is no vertex of τ_B strongly dominates J
- 2) $\exists J \in V - \tau_B$; is the only vertex in τ_B which strongly dominates J
- 3) τ_B is bipolar intuitionistic fuzzy graph with total dominating set.

Proof: Suppose that τ_B is a minimal total strong (weak) dominating set. Then for every $J \in \tau_B, \tau_B - \{J\}$ is not a total strong (weak) dominating set, then there exist $I \in V - \tau_B$, such that not strongly dominated by any vertex belong to $\tau_B - \{J\}$. Since τ_B is total strong dominating set. Thus J is only vertex which strongly dominates I and hence axiom 2 hold.

Now, suppose that τ_B be total strong (weak) dominating set also for each $J \in \tau_B$ thus one of the following two axioms holds

If τ_B not minimal, then $\exists J \in \tau_B, \tau_B - \{J\}$ a total strong dominating set also hence J is strongly dominated by at least one vertex in $\tau_B - \{J\}$ and its contradiction by 1

If $\tau_B - \{J\}$ is a total strong (weak) dominating, then every vertex belong to $V - \tau_B$ is totally strong (weak) dominated by at least one vertex belong to $\tau_B - \{J\}$. The second condition does not holds. This τ_B a minimal total strong (weak) dominating set.

Corollary 3.17: All complete bipolar intuitionistic fuzzy graph also total strong (weak) domination in bipolar intuitionistic fuzzy graphs.

Proof: G be complete bipolar intuitionistic fuzzy graphs. Thus every edge is total strong (weak) dominating set also every vertices are join together. Thus and obvious, G be total strong (weak) dominating set.

Example 3.18: Given a bipolar intuitionistic fuzzy graphs G given by Fig. 2 and 3.

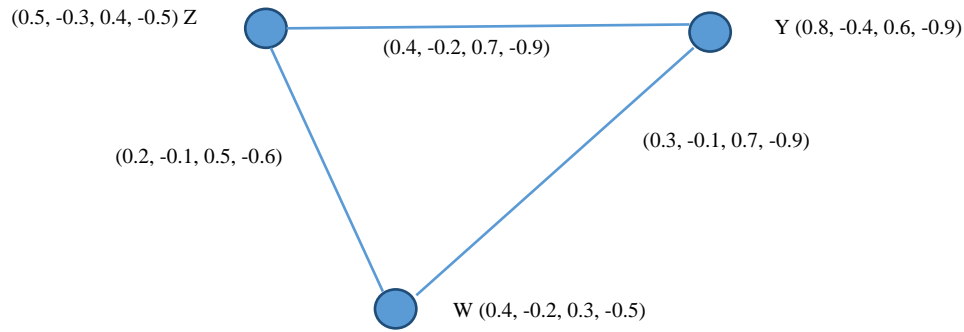


Fig. 2. Bipolar Intuitionistic Fuzzy Graph.

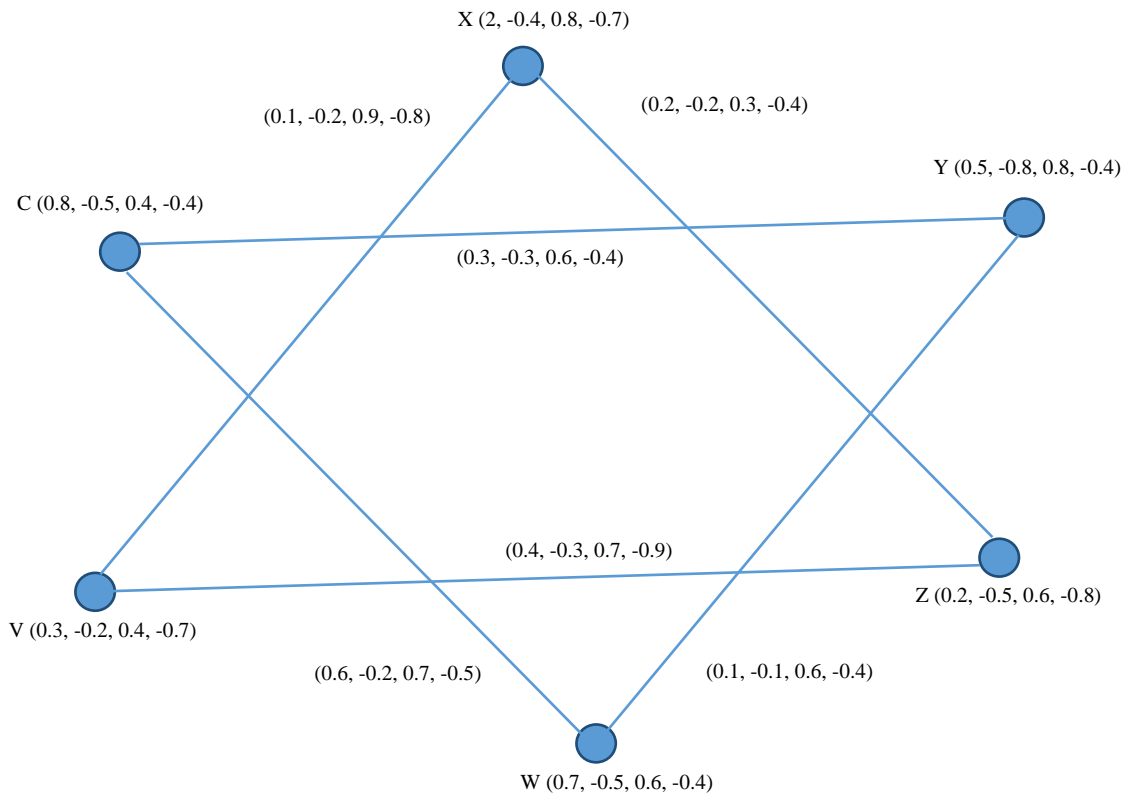


Fig. 3. Bipolar Intuitionistic Fuzzy Graph.

$$\tau_B - \{Z\}, \pi_{\tau_B} - \{G\} = (0.2, -0.1, 0.5, -0.6)$$

$$\tau_B - \{W\}, \pi_{\tau_B} - \{G\} = (0.3, -0.9, 0.4, -0.8)$$

Definition 3.19: A bipolar intuitionistic fuzzy graphs G said to be a semi - Ψ - strong bipolar intuitionistic fuzzy graph when $\mu_{\tau_{32}}^+(e, f) = \min\{\mu_{\tau_{31}}^+(e), \mu_{\tau_{31}}^+(f)\}$, and $\mu_{\tau_{32}}^-(e, f) = \min\{\mu_{\tau_{31}}^-(e), \mu_{\tau_{31}}^-(f)\}$ for every e and f .

Theorem 3.20: Let G be a bipolar intuitionistic fuzzy graph on order $O(G)$ then

$$1) \mathfrak{F}_{sbif}(G) \leq \mathfrak{F}_{tsbif}(G) \leq O(G) - \Delta_n(G) \leq O(G) - \Delta_e(G)$$

$$2) \mathfrak{F}_{sbif}(G) \leq \mathfrak{F}_{twbif}(G) \leq O(G) - \delta_n(G) \leq O(G) - \delta_e(G)$$

Where \mathfrak{F}_{sbif} , \mathfrak{F}_{tsbif} , \mathfrak{F}_{twbif} represent to strong bipolar intuitionistic domination, total strong bipolar intuitionistic domination and total strong (weak) bipolar intuitionistic domination respectively.

Proof: Hence G every \mathfrak{F}_{sbifd} - set (\mathfrak{F}_{twbifd} - set) is a bipolar intuitionistic dominating set $\mathfrak{F}_{sbif}(G) \leq \mathfrak{F}_{tsbif}(G)$ and $\mathfrak{F}_{wbif}(G) \leq \mathfrak{F}_{twbif}(G)$ if $i, j \in V$ and $D_n(i) = \Delta_n(G)$ and $D_n(i) = \delta_n(G)$ then $V - N(i)$ is a \mathfrak{F}_{tsbifd} - set but not minimal and $V - N(j)$ is \mathfrak{F}_{twbifd} - set but not minimal

hence $\mathfrak{J}_{tsbif}(G) \leq |V - N(i)|_{sbif}$ this means that $|V - N(i)|_{sbif} = |V| - |N(u)|$

$$\Rightarrow O(G) - D_n(u)$$

$$\Rightarrow O(G) - \Delta_n(G)$$

$$\Rightarrow \mathfrak{J}_{tsbif}(G) \leq O(G) - \Delta_n(G) \text{ and } \mathfrak{J}_{twbif}(G) \leq |V - N(i)|_{wbif}$$

this means that $|V - N(j)|_{wbif} = |V| - |N(j)|$

$$\Rightarrow O(G) - D_n(j)$$

$$\Rightarrow O(G) - \delta_n(G)$$

$$\Rightarrow \mathfrak{J}_{twbif}(G) \leq O(G) - \delta_n(G)$$

More over, $\Delta_e(G) \leq O(G)$ and $\delta_e(G) \leq \delta_n(G)$

$$\Rightarrow \mathfrak{J}_{sbif}(G) \leq \mathfrak{J}_{tsbif}(G) \leq O(G) - \Delta_n(G) \leq O(G) - \Delta_e(G)$$

$$\Rightarrow \mathfrak{J}_{wbif}(G) \leq \mathfrak{J}_{twbif}(G) \leq O(G) - \delta_n(G) \leq O(G) - \delta_e(G).$$

Corollary 3.21: If G is a bipolar intuitionistic fuzzy graph, then $\mathfrak{J}_{tsbif}(G) \leq \mathfrak{J}_{twbif}(G)$.

Proof: If x, y is a minimal total and strong weak dominating set respectively. If $D_n(i) = \Delta_n(G)$ and $D_n(j) = \delta_n(G)$ not that $V - N(j)$ is a total weak domination.

Example 4.1

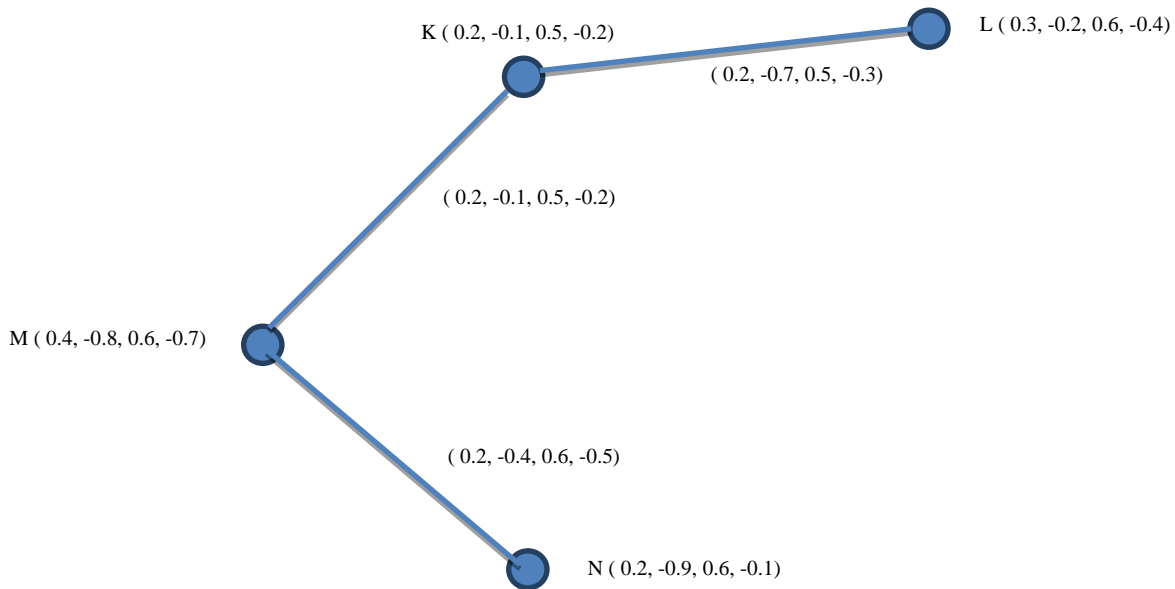


Fig. 4. Total Strong Domination Bipolar Intuitionistic Fuzzy Graphs.

$$\mathfrak{J}_{tsbif}(G) \leq |V - N(i)|_{sbif}$$

$$\mathfrak{J}_{tsbif}(G) \leq O(G) - \Delta_n(G) \text{ and } \mathfrak{J}_{twbif}(G) \leq |V - N(i)|_{wbif}$$

$$\mathfrak{J}_{twbif}(G) \leq O(G) - \delta_n(G)$$

Since $O(G) - \Delta_n(G) \leq O(G) - \delta_n(G)$ were we get

$$\mathfrak{J}_{Esbif}(G) \leq \mathfrak{J}_{Ewbif}(G)$$

Corollary 3.22: For a bipolar intuitionistic fuzzy graph G

- 1) $O(G) - S(G) \leq \mathfrak{J}_{Esbif}(G) \leq O(G) - \delta_e(G)$.
- 2) $O(G) - S(G) \leq \mathfrak{J}_{Ewbif}(G) \leq O(G) - \Delta_e(G)$.

Proof: Straight forward.

Note 3.23: Let G is a bipolar intuitionistic fuzzy graphs such that every vertex having a same membership grade, then

- 1) $O(G) - S(G) \leq \mathfrak{J}_{tsbif}(G) \leq O(G) - \delta_e(G)$.
- 2) $O(G) - S(G) \leq \mathfrak{J}_{twbif}(G) \leq O(G) - \Delta_e(G)$.

IV. EXAMPLES FOR TOTAL STRONG DOMINATION BIPOLAR INTUITIONISTIC FUZZY GRAPH

In this section we will provide examples of total strong domination bipolar intuitionistic fuzzy graphs illustrated by figures (Fig. 4, Fig. 5 and Fig. 6), respectively.

Example 4.2

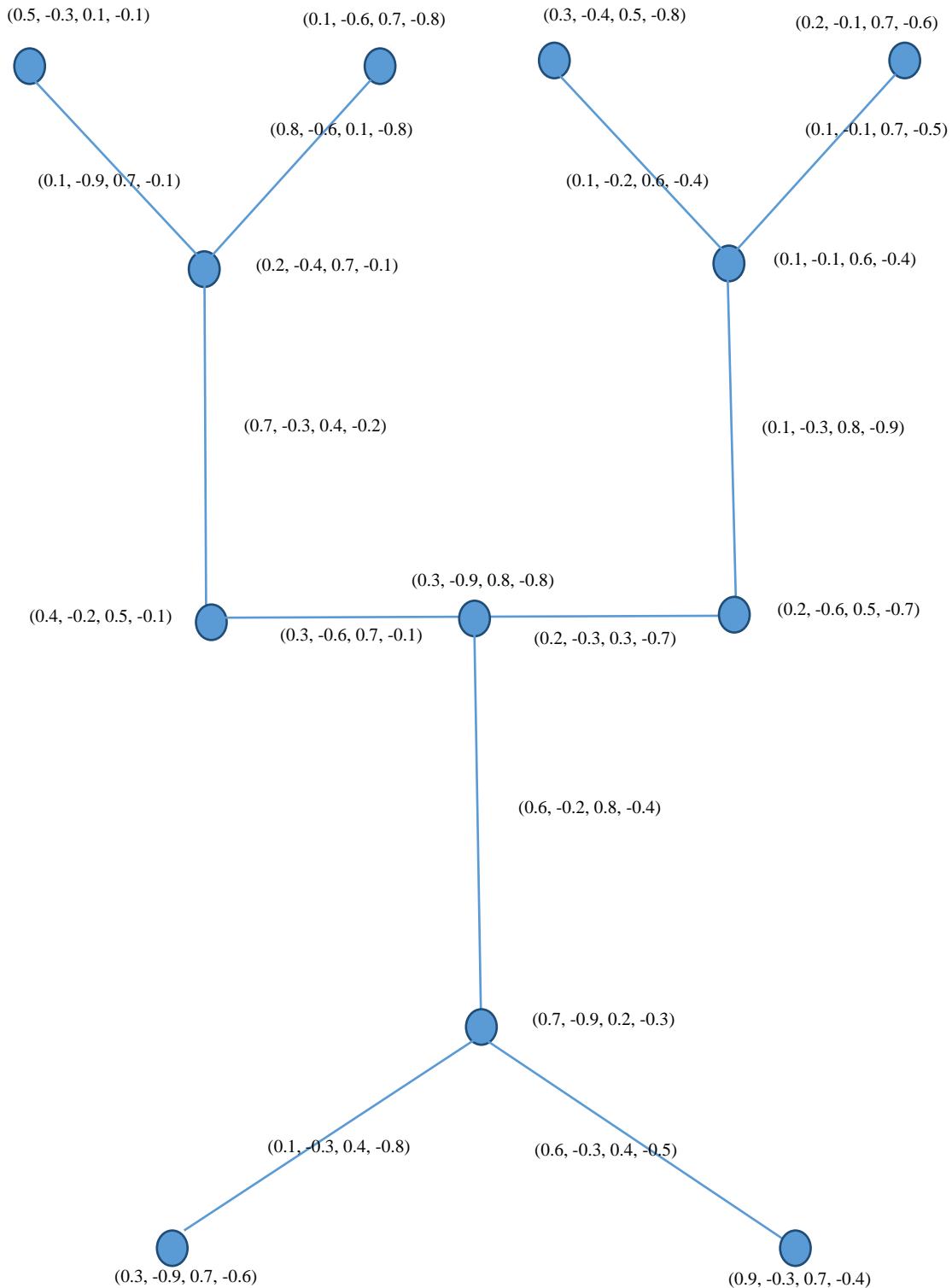


Fig. 5. Total Strong Domination Bipolar Intuitionistic Fuzzy Graphs.

Example 4.3

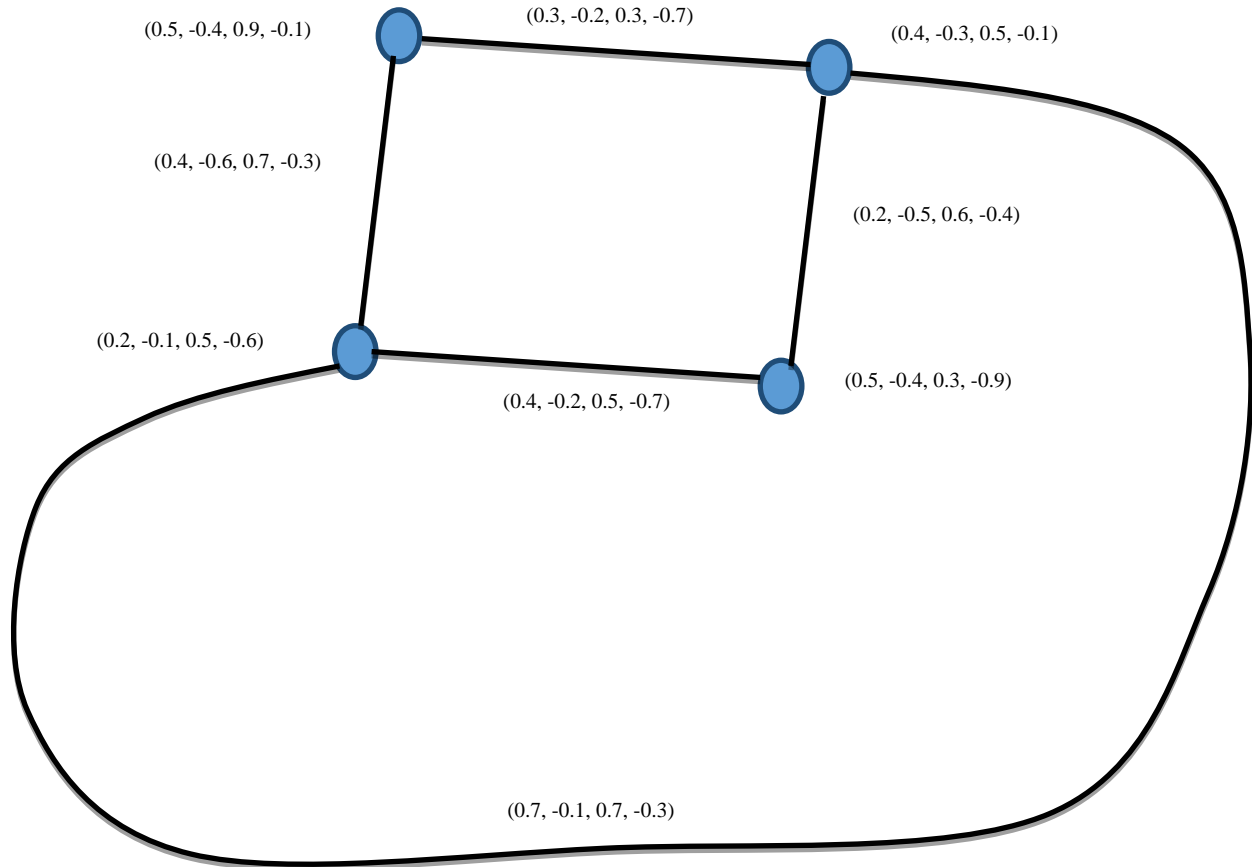


Fig. 6. Total Strong Domination Bipolar Intuitionistic Fuzzy Graphs.

V. CONCLUSION

Through this article, total strong (weak) domination, strong (weak) bipolar intuitionistic fuzzy graphs and some properties are discussed. In Over future paper these concepts will be generalized to hyper and tripolar fuzzy soft set, also this concept will be used in developing some topics in many papers such as [25- 27].

ACKNOWLEDGMENTS

The author is very grateful to the reviewers who will give their opinion on this manuscript. He also thank to their management for their constant encouragement in scientific research.

REFERENCES

[1] L. A. Zadeh, "Fuzzy set," *Information control*, vol. 8, 1965, pp. 338–353.
 [2] A. Rosenfeld, "Fuzzy groups," *Journal of Mathematical Analysis and Application*, vol. 35, no. 3, 1971, pp. 512-517.
 [3] A. Rosenfeld, *Fuzzy Graphs*, in L. A. Zadeh, Fu.k.s. shimura (E)s., "Fuzzy set and their application to cognitive and decision processes," *Academic press*, New York, 1975, pp. 77–95.
 [4] R.T. Yeh, S.Y. Bany, *Fuzzy relations and their applications to clustering analysis*, in L.A.Zadeh, k.s. Fu.M. shimura (E)s. *Fuzzy set and their application*, *Academic press*, 1975, pp. 125–149.
 [5] M.O.Massa'deh , Y.Al-wadi and F. Esma'el, "The fuzzy index and it's types," *Far East J. Math. Sci*, vol. 25, 2007, pp. 83-92.

[6] M.O.Massa'deh, "Some properties of upper fuzzy order, *African journal of mathematics and computer science research*," vol. 3, 2010, pp. 192-194.
 [7] M.O.Massa'deh, "On M-fuzzy subrings," *Far East journal of mathematical sciences*, vol. 62, 2011, pp. 41-49.
 [8] R.M.S.Mahmood and M.O.Massa'deh, "on Groups Acting on trees of finite extensions of free groups ," *Mathematical Sciences letters*, vol. 7 , 2018, pp. 111-116.
 [9] K.T. Atanassov, "intuitionistic fuzzy sets," *fuzzy sets and systems*, vol. 20, 1986, pp. 87-96.
 [10] M. Akram and B. Davvaz," Strong intuitionistic fuzzy graphs," *Filomat*, vol. 26, 2012, pp. 177-196.
 [11] M.O.Massa'deh, "Structure properties of an intuitionistic anti fuzzy subgraphs," *Journal of Applied Computer Science and Mathematics*, vol. 7, 2013, pp. 42-44.
 [12] M.O.Massa'deh, "A study on intuitionistic fuzzy and normal fuzzy M-subgroup, M-Homomorphism and Isomorphism ," *International Journal of industrial mathematics*, vol. 8, 2015, pp. 185-188.
 [13] M.O.Massa'deh," Some contributions on intuitionistic Q-fuzzy ku-Ideals," *JP Journal of Algebra, Number theory and Applications*, vol. 42, 2019, pp. 95-110.
 [14] W.R. Zhang, "Bipolar fuzzy sets," *In proceeding of Fuzz-IEEE*, 1998 , pp. 835-840.
 [15] M. Sunil, M.S. Sunitha and N. Anjali, "Some connectivity concepts in bipolar fuzzy graphs," *Annals of pure and applied mathematics*, vol. 7, 2014, pp. 98-108.
 [16] M.O.Massa'deh, "On bipolar fuzzy cosets," *bipolar fuzzy ideals and isomorphism of Γ- near rings*,vol. 102, 2017, pp. 731-747.

- [17] M.O.Massa'deh, "A study on anti bipolar Q-fuzzy normal semi groups," *Journal of Mathematical sciences and applications* , vol. 6, 2018, pp. 1-5.
- [18] W.A.Al-Zoubi , A.M.Alnaser, H.Hatamleh, Y.Alwadi and M.O.Massa'deh, "Introduction to Cartesian, tensor and lexi cographic product of bipolar interval valued fuzzy graph ," *Journal of Engineering and Applied Sciences*, vol. 15, 2018, pp. 581-585.
- [19] K. Sankar and D. Ezhilmaran, "bipolar intuitionistic fuzzy graphs with application," *International Journal of Research and Scientifical Innovation (IJRSI)*, vol. 3, 2016, pp. 44-52.
- [20] D. Ezhilmaran and K. Sankar, " morphism of bipolar intuitionistic fuzzy graphs," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 18, 2015, pp. 605 – 621.
- [21] A.M. Alnaser, W.A. Al-Zoubi and M.O.Massa'deh, "Bipolar intuitionistic fuzzy graph and it's matrices," *Applied Mathematics and information sciences*, vol. 14, 2020, pp. 205-214.
- [22] M. Akram and W.A. Dudak, " Regular bipolar fuzzy graphs", *Neural Comput and Applic*, vol.21, 2012, pp. 197-205.
- [23] M. B. Sheeba and R. Pilakkat, " Strength of Fuzzy cycles", *South Asian Journal of Mathematics*, vol. 3, 2013, pp. 8 – 12.
- [24] F. Harary, "Graph theory," Third edition, Addison-wesly, Reading , MA,1972.
- [25] A. M. Alnaser, Y. O. Kulakov, "Reliable Multipath Secure Routing In Mobile Computer Networks," *Computer Engineering and Intelligent Systems*, vol. 4, 2013, pp. 8-15.
- [26] A. M. Alnaser, "Set-theoretic Foundations of the Modern Relational Databases: Representations of Table Algebras Operations," *British Journal of Mathematics & Computer Science*, vol. 4, 2014, pp. 3286-3293.
- [27] A. M. Alnaser, "Streaming Algorithm For Multi-path Secure Routing in Mobile Networks," *IJCSI International Journal of Computer Science Issues*, vol. 11, 2014, pp. 112-114.

Performance Comparison of Three Hybridization Categories to Solve Multi-objective Flow Shop Scheduling Problem

A Case Study from the Automotive Industry

Jebari Hakim^{1*}, Siham Rekiek², Rahali El Azzouzi Saida³, Samadi Hassan⁴

Department of Information and Communication Systems
National School of Applied Sciences, University Abdelmalek Essaâdi
Tangier, Morocco

Abstract—The industries must preserve a rate of constant productivity; however, weaknesses appear at the level of production system which engenders high manufacturing costs. Scheduling is considered the most significant issue in the production system, the solution to that problem need complex methods to solve it. The goal of this paper is to establish three hybridization categories of the evolutionary methods ABC and PSO to solve multi-objective flow shop scheduling problem: Synchronous parallel hybridization using the weighted sum method of the fitness function, sequential hybridization using or not using the weighted sum method of the fitness function, and asynchronous parallel hybridization using the weighted sum method of the fitness function, then to test these methods in an automotive multi-objective flow shop and to perform an in-depth comparison for verifying how the multi hybridization and the hybridization categories influence the resolution of multi-objective flow shop scheduling problems. The results are consistent with other studies that have shown that the multi hybridization improves the effectiveness of the algorithm.

Keywords—Scheduling; multi-objective; flow shop; multi hybridization; artificial bee colony ABC; particle swarm optimization PSO

I. INTRODUCTION

The objectives of companies are diversified and the scheduling became multi-criterion. The scheduling objective are related to the time or the resources or the cost.

The scheduling problem in the production system is a accomplishment of a tasks group by taking in consideration some constraints.

The hybrid metaheuristics are proposed by Talbi [1] and are classified in three classification [2]:

- Synchronous parallel hybridization consists of incorporating an approach in an operator of another approach.
- Sequential hybridization is composed by various approaches, the solution of the first approach is an initialization of the next approach.

- Asynchronous parallel hybridization, the hybrid approaches share data throughout the search process.

In the flow shop scheduling problem, every machine can make only a single operation simultaneously and every job can have just a single operation in progress at the same instant. The capacity of storage inter-machines is defined and the preemption of operations is not approved.

Solving multi-objective flow shop scheduling problem has been gaining importance in recent years, in fact, many authors have developed diverse hybrid approaches and not hybrid approaches : Genetic local search [3], artificial neural network [4], particle swarm optimization [5], ant colony system [6], GRASP heuristic [7], hybrid TP+PLS [8], pareto approach [9], [10], [11], [12], multi-objective genetic algorithm and sub-population genetic algorithm-II and non-dominated sorting genetic algorithm-II [13], multi-objective genetic algorithm [14], quantum differential evolutionary algorithm [15], Parallel multiple reference point approach [16], glowworm swarm optimization [17], genetic algorithm [18], genetic algorithm optimization technique [19], memetic algorithm [20], hybrid non-dominated sorting genetic algorithm with variable local search [21], hybrid harmony search [22], Heuristic algorithms [23], lower-bound-based GA [24]. The author in [25] summarizes some contributions to solve flow shop scheduling problem.

However, to the authors' knowledge, very few publications are available in the literature that performed an in-depth comparison for verifying how the multi hybridization and the hybridization categories influence the resolution of multi-objective flow shop scheduling problem.

The objective of this paper is as follows:

- To establish three hybridization categories of the evolutionary methods ABC and PSO to solve multi-objective flow shop scheduling problem: Synchronous parallel hybridization using the weighted sum method of the fitness function, sequential hybridization using or not using the weighted sum method of the fitness function, and asynchronous parallel hybridization using the weighted sum method of the fitness function.

*corresponding Author

- To make tests of these methods in an automotive multi-objective flow shop.
- To perform an in-depth comparison for verifying how the multi hybridization and the hybridization categories influence the resolution of multi-objective flow shop scheduling problems.

The rest of the paper is organized as follows: the fundamentals of ABC and PSO will be explained in Section 2. In Section 3, the authors described the implementation of the proposed methods. The results and discussion are explained in Section 4. In Section 5, the conclusion and perspectives for further research are presented.

II. THE MATERIAL AND METHOD

A. Fundamentals of Artificial Bee Colony Algorithm

The artificial bee colony ABC algorithm is one of the most newly added swarm-based algorithms. ABC method created by Karaboga, it was copied the intelligent foraging behavior observed in the domestic bees to take the process of foraging [26].

ABC technique was produced for optimization problems in the continuous field. Recently, it was further enlarged for optimization problems in the discrete area [27] [28] [29] [30] [31].

A complete review of the utilization of ABC algorithm can be found in [32].

Four phases make ABC: Initialization bee phase, employed bees phase associate with particular food sources, onlooker bees phase look at the dance of engaged bees within the hive to choose a food source, and scout bees phase search randomly food sources.

In the ABC algorithm, the position of a food source corresponds to a possible feasible solution to the studied problem, and the nectar amount of a food source design the fitness of the solution.

The ABC algorithm merges techniques of local search and global search, trying to balance between the exploration and the exploitation of the search zone.

The main steps of the ABC method are as follows:

Initialization Phase (Initialize Population) REPEAT Employed Bees Phase (Put the employed bees on their food sources) Onlooker Bees Phase (Put the onlooker bees on the food sources according to their nectar amounts) Scout Bees Phase (Send the scouts to the search zone for exploring new food sources) Record the best food source attained so far UNTIL requirements are met

B. Fundamentals of Particle Swarm Optimization

Particle Swarm Optimization PSO technique is a popular swarm-intelligence-based algorithm that optimizes a problem

using an approach that is motivated by the movements of schooling of fishes or a flock of birds.

It was founded by Eberhart and Kennedy in 1995 [33] and has received significant attention from researchers studying in several research fields and has been successfully employed to many optimization problems since then [34] [35] [36].

The candidate solutions of a studied problem are designed as particles that form a population. The location of each particle is determined with two swarm main characteristics: the particle's position and velocity. The position of a particle represents a specific solution to the studied problem, while velocity is employed to define the direction of the particle in the next iteration.

Two reference values manage the movement of a particle throughout the iterations: The best fitness value obtained by the particle and the best fitness value of the swarm registered so far. The PSO has a memory that deposits the best fitness value of all particles achieved so far, and the corresponding position.

Applying these principles, improvement is accomplished and the PSO is conducted to the optimal solution.

The main steps of PSO are as follows:

Randomly generate the initial population and the velocities Repeat Determine the best values of particles in the swarm Change the best particles in the swarm Determine the best particle Update the velocities of particles Update the particle position Until requirements are met

C. The Proposed Methods

The ABC has a high capacity to explore the global optimum who it is not immediately employed, because the ABC stocks it at each iteration, on the other hand, the PSO can immediately employ the global best solution at iteration.

To obtain a better-performing method that exploits and combines advantages of these algorithms, the proposed hybridization between the ABC and PSO is applied.

The proposed hybrid metaheuristics are developed in three-hybridization categories: Synchronous parallel hybridization, sequential hybridization and asynchronous parallel hybridization.

1) *Synchronous parallel multiple hybridization of ABC with PSO*: The authors developed a new approach of synchronous parallel hybridization of ABC and PSO called HABCPSO. This approach consists to employ in the employed bees phase or/and in the onlooker bees phase or/and in the scout bees phase, the position and the velocity updating process, the Table I shown the configuration of the HABCPSO methods.

Consequently, the procedure of HABCPSO2 and HABCPSO3 can be found in Fig. 1 and Fig. 2, respectively.

The fitness function F minimized in HABCPSO corresponds to the balanced sum of both objectives functions F_1 and F_2 , with weights β_1 and β_2 defined as follows:

$$F = F_1 \beta_1 + F_2 \beta_2, \beta_1 + \beta_2 = 1, \beta_1 > 0, \beta_2 > 0 \quad (1)$$

1) *Sequential hybridization of ABC with PSO*: The proposed hybrid methods, denoted as [ABC+PSO](F) and ABC(F1)+PSO(F2), are founded on the recombination of two procedures ABC and PSO. The PSO is applied after the ABC.

PSO has used the output of the previous as its inputs, there are acting in a pipeline way.

The procedures of [ABC+PSO](F) and ABC(F1)+PSO(F2) are illustrated in Fig. 3.

The fitness function F minimized in [ABC+PSO](F) corresponds to the balanced sum of both objectives functions F_1 and F_2 , with weights β_1 and β_2 defined by the function 1.

The fitness functions F_1/F_2 minimized in ABC(F1)+PSO(F2) corresponds to objective function F_1 minimized in ABC(F1) and objective function F_2 minimized in PSO(F2).

2) *Sequential hybridization of PSO with ABC*: The proposed hybrid methods denoted as [PSO+ABC](F) and PSO(F1)+ABC(F2) are founded on the recombination of two procedures ABC and PSO. The ABC is applied after the PSO.

ABC has used the output of the previous as its inputs, there are acting in a pipeline way.

The procedures of [PSO+ABC](F) and PSO(F1)+ABC(F2) are illustrated in Fig. 4.

TABLE I. THE CONFIGURATION OF THE HABCPSO METHODS

Hybrid ABC + PSO	ABC		
	Employed bee phase	Onlooker bee phase	Scout bee phase
HABCPSO1	position and velocity updation process		
HABCPSO2	position and velocity updation process	position and velocity updation process	
HABCPSO3	position and velocity updation process	position and velocity updation process	position and velocity updation process
HABCPSO4		position and velocity updation process	
HABCPSO5		position and velocity updation process	position and velocity updation process
HABCPSO6	position and velocity updation process		position and velocity updation process
HABCPSO7			position and velocity updation process

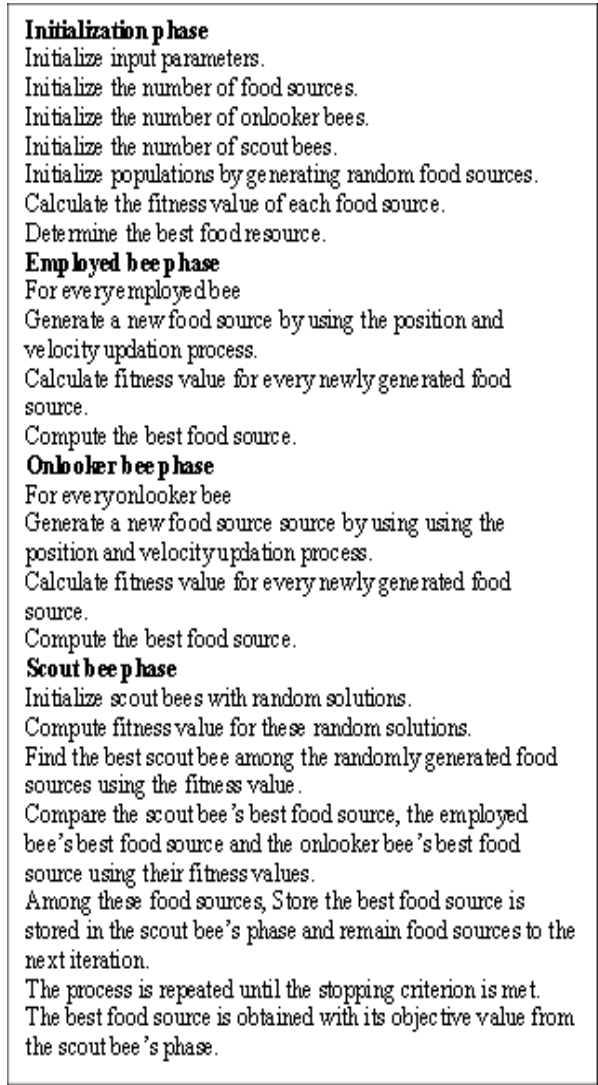


Fig. 1. The Procedure of HABCPSO2.

The fitness function F minimized in [PSO+ABC](F) corresponds to the balanced sum of both objectives functions F_1 and F_2 , with weights β_1 and β_2 defined by the function 1.

The fitness functions F_1/F_2 minimized in PSO(F1)+ABC(F2) corresponds to objective function F_1 minimized in PSO(F1) and objective function F_2 minimized in ABC(F2).

3) *Asynchronous parallel hybridization of ABC with PSO*: The proposed hybrid method denoted as ABC//PSO is founded on the recombination of two procedures ABC and PSO, this share and exchange information throughout the search process. The procedures of ABC//PSO algorithm are illustrated in Fig. 5.

The fitness function F minimized in ABC//PSO corresponds to the balanced sum of both objectives functions F_1 and F_2 , with weights β_1 and β_2 defined by the function 1.

Initialization phase
 Initialize input parameters.
 Initialize the number of food sources.
 Initialize the number of onlooker bees.
 Initialize the number of scout bees.
 Initialize populations by generating random food sources.
 Calculate the fitness value of each food source.
 Determine the best food resource.

Employed bee phase
 For every employed bee
 Generate a new food source by using the position and velocity updation process.
 Calculate fitness value for every newly generated food source.
 Compute the best food source.

Onlooker bee phase
 For every onlooker bee
 Generate a new food source by using the position and velocity updation process.
 Calculate fitness value for every newly generated food source.
 Compute the best food source.

Scout bee phase
 For every Scout bee
 Generate a new food source by using the position and velocity updation process.
 Calculate fitness value for every newly generated food source.
 Compute the best food source.
 Compare the scout bee's best food source, the employed bee's best food source and the onlooker bee's best food source using their fitness values.
 Among these food sources, Store the best food source is stored in the scout bee's phase and remain food sources to the next iteration.
 The process is repeated until the stopping criterion is met.
 The best food source is obtained with its objective value from the scout bee's phase.

Fig. 2. The Procedure of HABCPSO3.

Initialize the population of food sources.
 Repeat
 Put the employed bees on their food sources in the memory and updating feasible food source
 Put the onlooker bees on the food sources depending on their nectar amounts
 Transmit the scouts to the search zone in order to discovering new food sources
 Memorize the best solution attained
 Until stop criterion is met
 Initialize the PSO Population
 Add the best solution of ABC to PSO population
 Repeat
 Calculate fitness values of particles
 Modify the best particles in the swarm
 Choose the best particle
 Calculate the velocities of particles
 Update the particle positions
 Memorize the best solution attained
 Until stop criterion is met
 Return the best solution

Fig. 3. The Procedure of [ABC+PSO](F) and ABC(F1)+PSO(F2).

Initialize the PSO Population
 Repeat
 Calculate fitness values of particles
 Modify the best particles in the swarm
 Choose the best particle
 Calculate the velocities of particles
 Update the particle positions
 Memorize the best solution attained
 Until stop criterion is met
 Initialize the population of food sources.
 Add the best solution of PSO to ABC population
 Repeat
 Put the employed bees on their food sources in the memory and updating feasible food source
 Put the onlooker bees on the food sources depending on their nectar amounts
 Transmit the scouts to the search zone in order to discovering new food sources |
 Memorize the best solution attained
 Until stop criterion is met
 Return the best solution

Fig. 4. The Procedure of [PSO+ABC](F) and PSO(F1)+ABC(F2).

Initialize the PSO Population
 Repeat
 Calculate fitness values of particles
 Modify the best particles in the swarm
 Choose the best particle
 Calculate the velocities of particles
 Update the particle positions
 Put the employed bees (best particles) on their food sources in the memory and updating feasible food source
 Put the onlooker bees on the food sources depending on their nectar amounts
 Transmit the scouts to the search zone in order to discovering new food sources
 Memorize the best solution attained
 Until stop criterion is met
 Return the best solution

Fig. 5. The Procedure of ABC//PSO.

III. RESULTS AND DISCUSSION

Authors are suggested to solve the multi-objective scheduling problem in the automotive company. This company produces automotive parts in different elastomeric materials, including silicone and TPE.

The automotive company workshop is a flow shop contains 17 production lines, each line is composed with seven workstations:

- M1: The injection machine.
- M2: The deburring workstation.
- M3: The inspection workstation.
- M4: The assembly workstation number 1.
- M5: The assembly workstation number 2.
- M6: The color control machine.
- M7: The inspection workstation of the finished product.

The real result of scheduling obtained from the production planner is shown in Table II.

Several operations of cleaning and tools change or parameter adjustment are managed in the workstations, parallelly with the production operations.

Dates are calculated from an initial time t_0 , and the time unit is expressed in minute.

The fitness function of total production cost is denoted F_1 and the fitness function of the stopping cost and the cost of not-use of the production line are denoted F_2 .

The fitness function F_1 and F_2 are given as follows:

$$F_1 = C_{prod}^{tot} = \sum_k \sum_i W_{ik} P_{ik} C_k^{ui} \quad (2)$$

$$W_{ik} = \begin{cases} 1: & \text{If the product is made in the production line} \\ 0: & \text{Otherwise} \end{cases}$$

$$F_2 = C_{arr}^{tot} = \sum_k C_{arr\ k} \sum_i W_{ik} tp_{ik}^{arr} + tp_{ik}^{nu} \text{ tel que } tp_{ik}^{arr} = D_{ik}^{nett} + D_{ik}^{chf}, \quad (3)$$

$$W_{ik} = \begin{cases} 1: & \text{If the product is made in the production line} \\ 0: & \text{Otherwise} \end{cases}$$

Consequently, the fitness function F is given as follows:

$$F = F_1 \beta_1 + F_2 \beta_2, \quad \beta_1 + \beta_2 = 1, \beta_1 > 0, \beta_2 > 0 \quad (4)$$

$$F = \beta_1 C_{prod}^{tot} + \beta_2 C_{arr}^{tot}, \beta_1 + \beta_2 = 1, \beta_1 > 0, \beta_2 > 0 \quad (5)$$

$$F = \beta_1 (\sum_k \sum_i W_{ik} P_{ik} C_k^{ui}) + \beta_2 (\sum_k C_{arr\ k} \sum_i W_{ik} tp_{ik}^{arr} + tp_{ik}^{nu}), \quad (6)$$

$$\beta_1 + \beta_2 = 1, \beta_1 > 0, \beta_2 > 0$$

The value considered for factors β_1 and β_2 is 0.5.

- F_{ik} : Manufacturing operation of the product i in the production line Ch_k .
- P_i : Finished product after the operation F_{ik} .
- P_{ik} : Manufacturing time of the operation F_{ik} .
- CP_{ik} : Time of the end of the execution of P_i in the production line Ch_k .
- CP_{i}^{stk} : Storage cost by unit of time of the product P_i .
- tp_{ik} : Setup time of the production line Ch_k before the operation F_{ik} .
- tp_{ik}^{arr} : Stopping time during the operation F_{ik} in the line Ch_k .
- tp_{ik}^{nu} : Time of no use of the line Ch_k before the operation F_{ik} .
- C_{prod}^{tot} : Total production cost.
- C_k^{ui} : The production unit cost of the product i in the production line Ch_k .

- D_{ik}^{nett} : Operations duration of the cleaning in the production line Ch_k .
- D_{ik}^{chf} : Changes format duration in the production line Ch_k .
- $C_{arr\ k}$: Stopping costs and no use of the production line Ch_k per unit time.
- C_{arr}^{tot} : Total stopping cost and no use of line by time unit.
- $trop_Ch_i$: The production time expressed in time unit.
- $tarr_Ch_i$: The stopping time expressed in time unit.
- $tnett_Ch_i$: The cleaning time expressed in time unit.
- $C_{pro_Ch_i}$: Production costs.
- $C_{nou_Ch_i}$: The costs of no use of the production line.

In the employed bees phase; the function of updated memory is as follows [22]:

$$y_{ij} = x_{ij} + \phi_{ij} (x_{ij} - x_{kj}), k \neq i, i = \{1, 2, \dots, SN\}, j = \{1, 2, \dots, D\}, \phi_{ij} = Rand[-1, 1] \quad (7)$$

x_{min} , x_{max} are respectively the lower bound and the upper bound of the search scope and y_{ij} is new feasible dimension value of the food sources that is modified from its previous food sources value x_{ij} .

In the onlooker bees phase, the probability value related with the food source (p_i) is as follows [22]:

$$p_i = \frac{fit_i}{\sum_{k=1}^{SN} fit_k} \quad (8)$$

fit_i is the fitness value of the solution.

In the scout bees phase, the transmission function is defined as follows [21]:

$$x_i^j = x_{min}^j + rand[0,1](x_{max}^j - x_{min}^j) \quad (9)$$

Each iteration a particle's velocity and a particle's position are updated according to the equation:

$$V_{k+1} = \mu V[k] + C1 * rand() * (pbest[k] - current[k]) + C2 * rand() * (gbest[k] - current[k]), C1 + C2 = 1 \quad (10)$$

μ is the inertia factor and used to control intensification and diversification, $V[]$ is the particle velocity and $C1/C2$ are the apprenticeship factors.

The algorithms were programmed in Java and executed in Core™ i7 CPU with 2.5GHz and 8 Go de RAM.

The ABC stopping criterion defines the maximum number of cycles that a food source can keep without improvement.

The $tarr_Ch_i$ and $tnett_Ch_i$ values of product i in each production line are given as follows:

$$\forall i = \{1, 2, 3, \dots, 17\} \quad tarr_Ch_i(P_i) \in [5, 10]$$

$$\forall i = \{1, 2, 3, \dots, 17\} \quad tnett_Ch_i(P_i) \in [5, 15]$$

The $trop_Ch_i$ value of product i in each production line is given as follows:

TABLE II. THE GLOBAL RESULTS OF THE PROPOSED METHODS AND THE REAL RESULTS

REAL DATA	Synchronous parallel hybridization of ABC with PSO							Sequential hybridization of ABC with PSO				Asynchronous parallel hybridization of ABC with PSO	
	HABCPSO1	HABCPSO2	HABCPSO3	HABCPSO4	HABCPSO5	HABCPSO6	HABCPSO7	ABC(F1) +PSO(F2)	PSO(F1) +ABC(F2)	ABC(F1) +PSO(F2)	PSO(F1) +ABC(F2)	ABC//PSO	
Number of jobs	Makespan												
25	512	454	351	306	421	318	318	466	366	421	351	401	351
40	591	524	415	370	491	382	382	536	430	491	415	471	415
60	679	608	491	446	573	458	458	620	506	573	491	553	491
85	783	700	558	495	659	516	516	712	573	659	558	625	558
110	832	737	583	520	696	541	541	760	606	696	583	662	583
130	865	748	585	522	710	543	543	773	610	710	585	672	585
160	901	787	620	557	749	578	578	814	649	749	620	711	620
190	936	822	629	566	776	587	587	851	659	776	629	731	629
230	1016	878	677	600	829	621	621	910	709	829	677	781	677
270	1077	926	715	638	875	659	659	961	751	875	715	823	715
310	1103	942	718	633	889	662	662	982	757	889	718	833	718
350	1183	1015	771	677	957	709	709	1057	812	957	771	895	771
400	1246	1071	813	719	1013	751	751	1115	865	1013	813	948	813
450	1294	1105	830	729	1047	768	768	1151	882	1047	830	978	830
500	1325	1135	851	750	1073	789	789	1185	903	1073	851	999	851
600	1482	1267	941	821	1205	870	870	1320	1003	1205	941	1126	941
700	1578	1341	995	869	1272	920	920	1396	1058	1272	995	1188	995
800	1686	1428	1058	913	1353	976	976	1488	1126	1353	1058	1262	1058
900	1795	1516	1118	945	1436	1023	1023	1580	1191	1436	1118	1336	1118

$$\forall i = \{1, 2, 3, \dots, 17\} \quad \text{trop_Ch}_i(P_i) \in [30,90]$$

The $C_{pro_Ch}_i$ and $C_{nou_Ch}_i$ values of product i in each production line are calculated according to formulas 2 and 3 and fitness function deduced from formulas 4 and 5.

The global results of the proposed methods are shown in Table II.

The computational results demonstrate that all proposed methods are given the best results compared with the real results in terms of solution quality.

The results show that the synchronous parallel hybridization method HABCPSO3 is given the best results compared with other results obtained by the sequential hybridization methods and the asynchronous parallel hybridization methods.

The results show that the synchronous parallel hybridization method HABCPSO3 is given the best results compared with other results obtained by the synchronous parallel hybridization methods, the synchronous parallel hybridization method HABCPSO5 is given the equal results to the results obtained by the synchronous parallel hybridization

methods HABCPSO6 independent of the hybridization type. These are presented in Fig. 6.

The ranking of the synchronous parallel hybridization methods in terms of performance according to the hybridization type is shown in Table III.

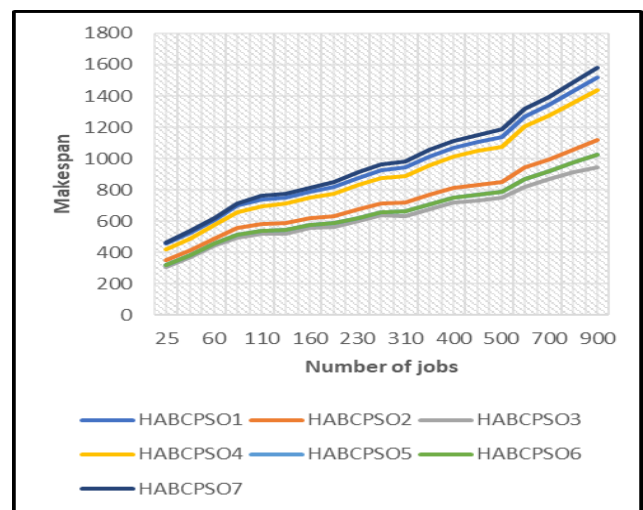


Fig. 6. The Results of the Synchronous Parallel Hybridization Methods.

TABLE III. THE RANKING OF THE SYNCHRONOUS PARALLEL HYBRIDIZATION METHODS

Ranking	Synchronous parallel hybridization of ABC and PSO			
	HABCPS OX	Employed bees phase	Onlooker bees phase	Scout bees phase
5	HABCPS O1	PSO		
3	HABCPS O2	PSO	PSO	
1	HABCPS O3	PSO	PSO	PSO
4	HABCPS O4		PSO	
2	HABCPS O5		PSO	PSO
2	HABCPS O6	PSO		PSO
6	HABCPS O7			PSO

Fig. 7 is shown that the sequential hybridization method ABC(F1)+PSO(F2) is given the best results compared with other results obtained by the other sequential hybridization methods.

As shown in Fig. 7:

- The sequential hybridization method [ABC+PSO](F) is given the best results compared with the results obtained by the sequential hybridization method [PSO+ABC](F).
- The sequential hybridization method ABC(F1)+PSO(F2) is given the best results compared with the results obtained by the sequential hybridization method PSO(F1)+ABC(F2).

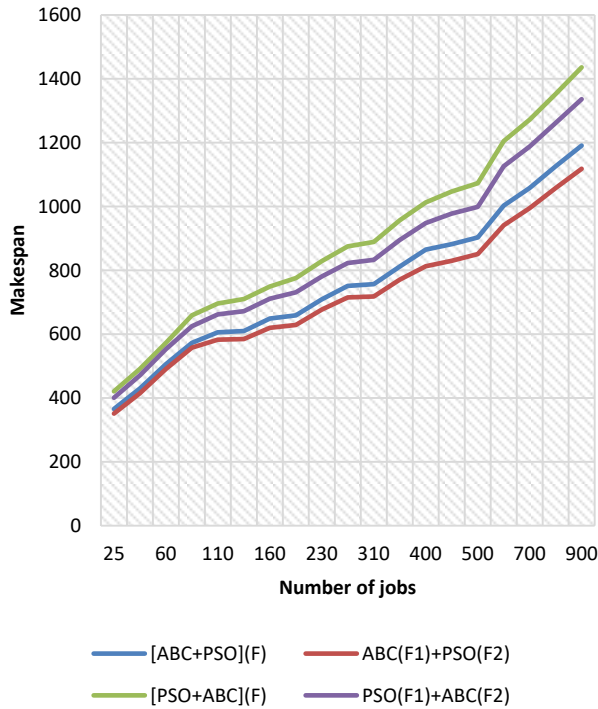


Fig. 7. The Results of the Sequential Hybridization Methods.

TABLE IV. THE RANKING OF THE SEQUENTIAL HYBRIDIZATION METHODS

Ranking	Sequential hybridization of ABC and PSO
1	ABC(F1)+PSO(F2)
2	[ABC+PSO](F)
3	PSO(F1)+ABC(F2)
4	[PSO+ABC](F)

The ranking of the sequential hybridization methods in terms of performance according to the fitness function F or F1 and F2 is shown in Table IV.

The results show that the asynchronous parallel hybridization method ABC//PSO is given the equal results to the results obtained by the sequential hybridization method ABC(F1)+PSO(F2), the asynchronous parallel hybridization method ABC//PSO is given the best results compared with other results obtained by the other sequential hybridization methods [ABC+PSO](F), PSO(F1)+ABC(F2), [PSO+ABC](F).

The ranking of the asynchronous parallel hybridization method and the sequential hybridization methods in terms of performance according to the fitness function F or F1 and F2 is shown in Table V.

As shown in Fig. 8:

- The synchronous parallel hybridization method HABCPSO2 is given the equal results to the results obtained by the sequential hybridization method ABC(F1)+PSO(F2) and the asynchronous parallel hybridization method ABC//PSO.
- The synchronous parallel hybridization HABCPSO4 is given the equal results to the results obtained by the Sequential hybridization [PSO+ABC](F).
- The synchronous parallel hybridization method HABCPSO2, the sequential hybridization method ABC(F1)+PSO(F2) and the asynchronous parallel hybridization method ABC//PSO are given the best results compared with other results obtained by the synchronous parallel hybridization HABCPSO4 and the sequential hybridization [PSO+ABC](F).
- The synchronous hybridization method ABC(F1)+PSO(F2) is given the best results compared with other results obtained by the synchronous hybridization method [PSO+ABC](F).

TABLE V. THE RANKING OF THE ASYNCHRONOUS PARALLEL HYBRIDIZATION METHOD AND THE SEQUENTIAL HYBRIDIZATION METHODS

Ranking	Hybrid ABC and PSO	
1	Asynchronous parallel hybridization of ABC and PSO	ABC//PSO
1	Sequential hybridization of ABC and PSO	ABC(F1)+PSO(F2)
2		[ABC+PSO](F)
3		PSO(F1)+ABC(F2)
4		[PSO+ABC](F)

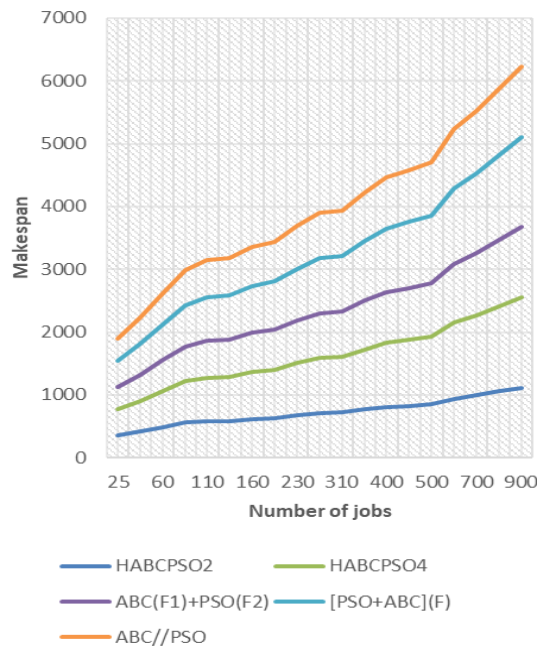


Fig. 8. The Results of the Synchronous Parallel Hybridization Methods HABCPSO2 / HABCPSO4 and the Asynchronous Parallel Hybridization Method ABC//PSO and the Sequential Hybridization Methods ABC(F1)+PSO(F2) / [PSO+ABC](F).

The ranking of the synchronous parallel hybridization methods HABCPSO2 / HABCPSO4 and the asynchronous parallel hybridization method ABC//PSO and the sequential hybridization methods ABC(F1)+PSO(F2) / [PSO+ABC](F) in terms of performance according to the fitness function is shown in Table VI.

The ranking of all proposed methods in terms of performance according to the fitness function F or F1 and F2 is shown in Table VII.

Summing up the results according to the ranking of the proposed methods in terms of performance, it can be concluded that:

- The proposed methods are given the best results compared with the real results of scheduling.
- The synchronous parallel hybridization of ABC in its three phases using the fitness function F is produced the best result.
- The synchronous parallel hybridization of ABC in its two phases (onlooker bees phase and scout bees phase) using the fitness function F is produced the equal results to the results of the synchronous parallel hybridization of ABC in its two phases (employed bees phase and

scout bees phase) using the weighted sum method of the fitness function F.

- The synchronous parallel hybridization of ABC in its two phases using the weighted sum method of the fitness function F (employed bees phase and onlooker bees phase) is given 100% results equal to the results of the sequential hybridization of ABC with PSO using the weighted sum method of the fitness function F and equal to the results of the asynchronous parallel hybridization of ABC with PSO using the weighted sum method of the fitness function F.
- The sequential hybridization of ABC with PSO using the weighted sum method of the fitness function F is given the better results than the sequential hybridization of PSO with ABC using the fitness functions F1 and F2.
- The synchronous parallel hybridization ABC in its onlooker bees phase using the weighted sum method of the fitness function F is produced the equal results to the results of the sequential hybridization of ABC with PSO using two fitness functions F1 and F2.
- The synchronous parallel hybridization of ABC in its employed bees phase using the weighted sum method of the fitness function F is produced the better result than the synchronous parallel hybridization of ABC in its scout bees phase using the weighted sum method of the fitness function F.

The authors' attention was concentrated not only on develops these three hybridization categories of the evolutionary methods ABC and PSO but also on tests of these hybrid methods in an automotive multi-objective flow shop and on their performance evaluation to make a perform comparison. The main limitation of the experimental that it is take into account the fitness function F1/F2 only in the sequential hybridization of ABC and PSO.

TABLE VI. THE SYNCHRONOUS PARALLEL HYBRIDIZATION METHODS HABCPSO2 / HABCPSO4 AND THE ASYNCHRONOUS PARALLEL HYBRIDIZATION METHOD ABC//PSO AND THE SEQUENTIAL HYBRIDIZATION METHODS ABC(F1)+PSO(F2) / [PSO+ABC](F)

Ranking	Hybrid ABC and PSO	
1	Asynchronous parallel hybridization of ABC and PSO	ABC//PSO
1	Sequential hybridization of ABC and PSO	ABC(F1)+PSO(F2)
2		[PSO+ABC](F)
1	Synchronous parallel hybridization of ABC and PSO	HABCPSO2
2		HABCPSO4

TABLE VII. THE RANKING OF ALL PROPOSED METHODS IN TERMS OF PERFORMANCE

Ranking	Hybrid ABC and PSO	Employed bees phase	Onlooker bees phase	Scout bees phase	Hybridation number	Fitness function	
1	Synchronous parallel hybridization of ABC and PSO	HABCPSO3	PSO	PSO	PSO	3	F
2	Synchronous parallel hybridization of ABC and PSO	HABCPSO5		PSO	PSO	2	F
2	Synchronous parallel hybridization of ABC and PSO	HABCPSO6	PSO		PSO	2	F
3	Synchronous parallel hybridization of ABC and PSO	HABCPSO2	PSO	PSO		2	F
3	Sequential hybridization of ABC and PSO	[PSO+ABC](F)				1	F
3	Asynchronous parallel hybridization of ABC and PSO	ABC//PSO				1	F
4	Sequential hybridization of ABC and PSO	[ABC+PSO](F)				1	F
5	Sequential hybridization of ABC and PSO	PSO(F1)+ABC(F2)				1	F1 and F2
6	Synchronous parallel hybridization of ABC and PSO	HABCPSO4		PSO		1	F
6	Sequential hybridization of ABC and PSO	ABC(F1)+PSO(F2)				1	F1 and F2
7	Synchronous parallel hybridization of ABC and PSO	HABCPSO1	PSO			1	F
8	Synchronous parallel hybridization of ABC and PSO	HABCPSO7			PSO	1	F

IV. CONCLUSION

Powerful methods to solve the multi-objective flow shop scheduling problem are required, due to high level of its complexity.

An adequate hybridization of multiple algorithmic concepts is the key to accomplishing top performance in solving scheduling problems.

Based on the overall experimental results, it can be decided that the proposed methods were capable to solve multi-objective flow shop scheduling problem successfully, efficiently, and robustly in terms of solution quality.

The paper presents a pilot study for verifying how the multi hybridization and the hybridization categories influence the resolution of multi-objective flow shop scheduling problems.

The proposed methods have great potential for other applications such as multi-objective job shop scheduling problem resolution and multi-objective open-shop scheduling problem resolution.

As a future research, we intend to apply the ideas presented in this paper to other scheduling problems such as multi-objective job shop scheduling problem and multi-objective open-shop scheduling problem using other hybrid methods in three hybridization categories.

REFERENCES

[1] E.G. Talbi, "A taxonomy of hybrid metaheuristics," International Journal of Heuristics, vol. 8, No. 5, pp. 541-564, 2002.

[2] D. Duvidier, "Etude de l'hybridation des méta-heuristiques, application à un problème d'ordonnement de type jobshop," These de Doctorat, Université du littoral France, France, Déc, 2000.

[3] J.E.C. Arroyo and V.A. Armentano, "Genetic local search for multi-objective flowshop scheduling problems," European Journal of Operational Research, vol. 167, No.3, pp. 717-738, 2005.

[4] A. Noorul Haq and T. Radha Ramanan, "A bicriterion flow shops scheduling using artificial neural network," The International Journal of Advanced Manufacturing Technology, vol. 30, No. 11-12, 2006.

[5] R. Rahimi-Vahed and S.M. Mirghorbani, "A multi-objective particle swarm for a flow shop scheduling problem," Journal of Combinatorial Optimization, vol. 13, No. 1, pp. 79-102, 2007.

[6] J.E.C. Arroyo and A.A. De Souza Pereira, "A GRASP heuristic for the multi-objective permutation flowshop scheduling problem," The International Journal of Advanced Manufacturing Technology, vol. 55, No. 5-8, pp. 741-753, 2010.

[7] J. Dubois-Lacoste, M. López-Ibáñez and T. Stützle, "A hybrid TP+PLS algorithm for bi-objective flow-shop scheduling problems," Computers & Operations Research, vol. 38, No. 8, pp. 1219-1236, 2011.

[8] M. Ciavotta, G. Minella and R. Ruiz, "Multi-objective sequence dependent setup times permutation flowshop: A new algorithm and a comprehensive study," European Journal of Operational Research, vol. 227, No. 2, pp. 301-313, 2013.

[9] Y. Collette and P. Siarry, "Optimisation Multiobjectif. Editions Eyrolles," Paris, 2002.

[10] Y. Sun, Ch. Zhang, L. Gao, and X. Wang, "Multi-objective optimization algorithms for flow shop scheduling problem: a review and prospects," International Journal of Advanced Manufacturing Technology, DOI 10.1007/s00170-010-3094-4, 2010.

[11] T. Loukil, J. Teghem and D. Tuytens, "Solving multi-objective production scheduling problems using metaheuristics," European Journal of Operational Research, vol. 161, No. 1, pp. 42-61, 2005.

[12] M. FADAEI, and M. ZANDIEH, "Scheduling a bi-objective hybrid flow shop with sequence-dependent family setup times using metaheuristics," Arabian Journal of Science Engineering, vol. 38, No. 8, pp. 2233-2244, 2013.

[13] F. Dugardin, F. Yalaoui, and L. Amodeo, "New multi-objective method to solve reentrant hybrid flow shop scheduling problem," European Journal of Operational Research, vol. 203, No. 1, pp. 22-31, DOI:10.1016/j.ejor.2009.06.031, 2010.

[14] B. Yagmahan and M.M. Yenisey, "A multi-objective ant colony system algorithm for flow shop scheduling problem," Expert Systems with Applications, vol. 37, No. 2, pp. 1361-1368, 2010.

[15] T.M. Zheng, M. and Yamashiro, "Solving flow shop scheduling problems by quantum differential evolutionary algorithm," The International Journal of Advanced Manufacturing Technology, vol. 49, No. 5-8, pp. 643-662, 2010.

[16] J.R. Figueira, A. Liefooghe, E.G. Talbi and A.P. Wierzbicki, "A parallel multiple reference point approach for multi-objective optimization," European Journal of Operational Research, vol. 205, No. 2, pp. 390-400, 2010.

[17] J. Senthilnath, S.N. Omkar, V. Mani, N. Tejovanth, P.G. Diwakar and S.B. Archana, "Multi-spectral satellite image classification using

- glowworm swarm optimization,” In IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 47-50, 2011.
- [18] P. Muni Babu, B.V. Himasekhar Sai and A. Sreenivasulu Reddy, “Optimization of make-span and total tardiness for flow-shop scheduling using genetic algorithm,” International Journal of Engineering Research and General Science, Vol. 3, Issue. 3, pp. 195-199, 2015.
- [19] G. Mohammadi, “Multi-Objective flow shop production scheduling via robust genetic algorithm optimization technique,” International Journal of Service Science, Management and Engineering, vol. 2, No. 1, pp. 1-8, 2015.
- [20] X. Wang and L. Tang, “A machine-learning based memetic algorithm for the multi-objective permutation flowshop scheduling problem,” Computers & Operations Research, vol. 79, Issue. C, pp. 60-77, 2017.
- [21] X. Wu, X. Shen and Q. Cui, “Multi-Objective Flexible Flow Shop Scheduling Problem Considering Variable Processing Time due to Renewable Energy,” Sustainability, vol. 10, No. 3, pp. 841, 2018.
- [22] Y. Li, X. Li, and J.N. Gupta, “Solving the multi-objective flowline manufacturing cell scheduling problem by hybrid harmony search,” Expert Systems with Applications, vol. 42, No. 3, pp. 1409-1417, 2015.
- [23] V. Arasanipalai Raghavan, S.W. Yoon and K. Srihari, “Heuristic algorithms to minimize total weighted tardiness with stochastic rework and reprocessing times,” Journal of Manufacturing Systems, vol. 37, pp. 233-242, DOI:10.1016/j.jmsy.2014.09.004, 2015.
- [24] A.J. Yu and J. Seif, “Minimizing tardiness and maintenance costs in flow shop scheduling by a lower-bound-based GA,” Computers & Industrial Engineering, vol. 97, pp. 26-40. DOI:10.1016/j.cie.2016.03.024, 2016.
- [25] T. El-Ghazali, F. Yalaoui and L. Amodio, “Metaheuristics for Production Systems 2015,” Springer 2015-11-26.
- [26] D. Karaboga and B. Basturk, “A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm,” Journal of Global Optimization, vol. 39, No. 3, pp. 459-471, 2007.
- [27] T.M. Pan Q-K, P. Suganthan and T. Chua, “A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem,” Inf Sci 181, pp. 2455-2468, 2011.
- [28] A. Singh, “An artificial bee colony algorithm for the leafconstrained minimum spanning tree problem,” Appl Soft Comput 9, pp. 625-631, 2009.
- [29] S. Sundar and A. Singh, “A swarm intelligence approach to the quadratic multiple knapsack problem,” In: ICONIP 2010. Lecture notes in computer science, vol 6443. Springer, Berlin, pp. 626-633, 2010.
- [30] T.M. Pan Q-K, P. Suganthan and AH-L. Chen, “A discrete artificial bee colony algorithm for the total flowtime minimization in permutation flow shops,” Inf Sci 181, pp. 3459-3475, 2011.
- [31] H. Jebari, S.R. Elazzouzi, H. Samadi and S. Rekiek, “The search of balance between diversification and intensification in artificial bee colony to solve job shop scheduling problem,” Journal of Theoretical and Applied Information Technology, vol. 97, No. 2, pp. 658-673, 2019.
- [32] D. Karaboga, B. Gorkemli, C. Ozturk and N. Karaboga, “A comprehensive survey: artificial bee colony (abc) algorithm and applications,” Artif Intell Rev 42(1), pp. 21-57, 2014.
- [33] M. Pontani and B.A. Conway, “Particle Swarm optimization applied to impulsive orbital transfers,” Acta Astronautica 74, pp. 141-155, May 2012.
- [34] R. Poli, “Analysis of the publications on the applications of particle swarm optimization,” Journal of Artificial Evolution and Applications, pp. 1-10, 2008.
- [35] K. Vaisakh, L.R. Srinivas and K. Meah, “Genetic evolving ant direction particle swarm optimization algorithm for optimal power flow with non-smooth cost functions and statistical analysis,” Appl. Soft Comput. 13(12), pp. 4579-4593, December 2013.
- [36] H. Jebari, S.R. Elazzouzi, H. Samadi and S. Rekiek, “Multi hybridization of swarm intelligence methods to solve job shop scheduling problem,” Journal of Theoretical and Applied Information Technology, vol. 97, No. 16, pp. 4366-4386, 2019.

Learners Classification for Personalized Learning Experience in e-Learning Systems

A. JOHN MARTIN¹

Research Scholar
Department of Computer Science
Sacred Heart College
Tirupattur, India

M. MARIA DOMINIC²

Assistant Professor
Department of Computer Science
Sacred Heart College
Tirupattur, India

F. Sagayaraj Francis³

Professor
Department of Computer Science
and Engineering, Puducherry
Technological, Pondicherry, India

Abstract—The investigators are inspired by the increasing need and the demand for educational applications and the Learning Management Systems which provide learning objects centered on the learning style of the learners. The technique in which the learners acquire, process, gain the information is unique; these unique characteristics affect their learning process. Hence it is essential to consider and understand the uniqueness among the learners to deliver learner-centric learning objects. The investigators present a system to classify the learners based on the time spent by the learner on learning content of different types. The types of learning content are identified with the percentage of visual, auditory, read/write and kinesthetic in learning object. The prominent learning style called VARK (Visual, Auditory, Read/Write and Kinesthetic) is used to classify the learners. This system classifies the learner and recommends the learning objects based on their learning preference, it also facilitates the faculty members or the content creators to prepare and provide personalized learning objects based on the learning style of the learners.

Keywords—Learning style; learning profile; learning objects; e-Learning; personalization

I. INTRODUCTION

Today, the need for education is the need of the hour in all the sectors. Learning can be defined as a change in the behaviour as a result of experience. The process of learning involves reception and transformation of received information. During the reception process diverse senses are engaged in gathering information from external sources, whereas transformation activity results in internal activities like memorization, inception, inference, pondering and reflection [14]. The acquisition of knowledge and processing the gained knowledge is uneven among learners. There are relative parameters that identify the learning style of a learner. Hence there is a need to adapt strategies to meet the learner preferences to in delivering the learning object whether they are physical or virtual. The process of personalization happens through the investigation of the student's preferences. It is possible to create a model that fulfils the need of the learner based on the information obtained through the investigation [1]-[3].

According to Bruner [3,15], the learner understands the knowledge through four sensory modes, they are Visual (screening pictures, symbols, chars and diagrams), Aural or

Auditory (listening, discussing with peer), Read / Write (reading and writing), and Kinesthetic (use of Hands on exercise, case studies, Demonstrations.). The learning style of a learner is determined by the way in which information is received and processed. Predefined mathematical equations and a set of questionnaires are the traditional ways used to determine learning style of the learners. This may not be appropriate because students prefer more than one mode of learning style because the percentage of time spent on each types of learning object will also vary. To satisfy a given learning style, the teacher or the content creator must use the approach that could meet the needs of diverse learning perspective. Hence the proposed system focuses on the following key contributions.

To recommend a novel but practical approaches to classify the learner based the time spent in each type of learning content to have personalized learning object or learner centred learning objects.

a) To experiment the work through an exemplary case with the data available in Arts and Science College.

The investigation is systematized as follows. Section II describes the State-of-Art of the existing system. Section III provides the proposed works that includes i. architectural design ii. The methodology to classify the learners based on the learner's preference and the experimental result and evaluation of the system have been explained in Section III. Finally the effectiveness of the system and future action plan is discussed.

II. STATE OF THE ART

A. Learning Objects

Any digital form or non-digital form of resource that are used to support learning activity is called Learning Objects. It is a collection of content items used by the learner in the technology assisted learning process. Instances of Learning Objects encompass multimedia content, reference to a web page, visuals, textual content, demonstration and software tools. The Learning objects will have the following characteristics size, duration, interoperability, reusability and multiple context of the content [13]. The significance of the learning object to the learner can be identified by the time spent on a particular learning content type.

B. Learning Styles

The learning style shows the way by which a personal collects, process, comprehends and retains the information is referred as learning styles. The learning styles rely upon emotional, cognitive and environmental factors, as well as prior experience of an individual.

There are several models for categorising the learning styles [11]. The popular learning models are discussed in the following sections.

1) David Kolb's learning style model

This learning style model has four stages, they are [12]

a) *Concrete Experience* – learning takes place through an exposure or circumstances encountered, or a through the modification of current exposure.

b) *Reflective Observation of the New Exposure* – learning takes place after gaining experience which enables learner to ask questions and discuss.

c) *Abstract Conceptualization* – enables the learner to get a new knowledge or a modification of an existing theoretical notion.

d) *Active Experimentation* - the learner tests their knowledge in the real world and gain new experience.

2) *Felder-silverman learning style model*: In accordance with Felder and Silverman, there are mixtures of components that make an impact on learning process; like visual/verbal, sensing/intuitive, sequential/global and active/reflective. As stated by Felder-Silverman the necessary teaching components or elements include visual/verbal, active/passive, sequential/global and concrete/abstract [8]. This model is most appropriate for the courses in engineering education [10].

3) *Honey and Mumford's learning styles*: As mentioned in the work of Kolb, Learning styles were evolved by Peter Honey and Alan Mumford. They identified four explicit learning styles: Activist, Theorist; Pragmatist and Reflector [11]. Honey and Mumford formed a set of questions that supports to identify individual's learning styles which is static in nature.

4) *Dunn and Dunn learning style model*: One of the oldest and most widely used approaches to learning styles is suggested by Rita and Kenneth Dunn (1978, 1992a, 1992b, and Dunn, 1986). According to Dunn and Dunn the learning style of learner classified into five dimensions [6].

- Environmental – The environment influences the learning style of a learner like sound, light, hotness and seating arrangements,
- Emotional – Related motivation, perseverance, and responsibility of the students.
- Sociological – This aspect identifies the preferences of the learning environment and how a learner prefers to learn in pair, what percentage of guidance is.

- Physiological – This is about how a learner responds to the learning task. It brings out other learning styles like visual, auditory and kinesthetic.
- Psychological – It is about how a learner process and respond to information and knowledge.

5) *VARK model*: In this investigation the researchers have chosen VARK model proposed by Neil D. Fleming [16]. It is one of the best model to classify the learning style. The attributes VARK (Visual, Auditory, Read/Write and Kinesthetic) constitutes Learning Object. The learning style and the learning approach of a learner based on VARK model is shown in the Table 1.

Understanding the individuals learning preferences can be helpful in the learning process. If a learner understands that visual learning suits the most preferred style, using visual study strategies in conjunction with other learning multimodal style might facilitate the learner to understand, interpret, remember and enjoy the learning.

C. Traditional Learning vs e-Learning

Though e-Learning is a full-fledged alternative for the classroom learning but it is not the substitute to the traditional learning. At the same time there exist good and ample evidence that the learner learns as much as online as they do in classroom learning. The major difference between traditional learning and e-Learning is, in traditional learning, learners are forced to study based on the syllabus irrespective of their likes and dislikes. In case of e-Learning, learner can filter and choose the content they want to learn. It also provides materials in various forms like audio, video, animations, presentations and documentation and so on.

The following Table 2 summarizes the difference between traditional and e-Learning [18].

TABLE I. SUMMARY OF VARK LEARNER MODEL

S.No	Learner Style	Learning approach towards learning	Learning Content
1	Visual	Acquisition of knowledge and understanding takes place through the images, maps, and graphic representations.	Video, URL, and Power points slides
2	Auditory	Learner understands content through listening, discussion and speaking	URL and Power Points slides with audio, recorded notes.
3	Read / Write	Prefer to learn from the text	Power Points Slides, Text documents and PDFs
4	Kinesthetic	Prefer to learn from project, practical, hands on experience, real time example. Learns by touch, feel, hold, move something.	Hands on exercise, case studies, Demonstrations.

TABLE II. THE DIFFERENCE BETWEEN TRADITIONAL AND E-LEARNING

S.No	Dimension	Traditional Learning	e-Learning
1	Discussion	Teacher discusses more than the student	Student discusses more than the teacher
2	Learning Process	For whole class, limited or no individual study	Learning process takes place with peer or an individual
3	Learning Objects	Decided by the teacher according to the curriculum	Student decides the learning object formats based on the learning preferences.
4	Emphasis on Learning process	Students learn "What" and not "how". Teachers are busy with completing the syllabus.	Students learn "how" and less "What"
5	Teacher's Role	Authority	Directs student to the information
6	Class Control	Control over the Learning object and presentation	Personalization on the learning object and presentation

D. Identification of Learning Styles

The memorization and processing of information by an individual is known as learning style [17]. There are numerous learning styles models, each offer diverse representation and classification with the types of learning. Every individual would have a unique learning style which is a significant attribute to provide personalized learning environment and to accomplish learning satisfaction [4,7,9]. Hence researchers have turned up to categorize the learners based on their learning style [19] since 1940s.

A survey was conducted to the engineering students; questions were based on Index of Learning styles proposed by Felder-Silverman, it was conducted in the year 2016 on 175 students studying MS programming in engineering [12]. A research had been conducted for the classification of learners' style with the VARK questionnaire from a sample group. [2]

Several studies have been done in the field of identification or prediction of learning styles majority of the studies have used survey methods where set of questionnaires were framed and the outcome of the survey shown as a result. Questionnaires were framed based on VARK learning style proposed by Flemming and Miles (1992) the same was used to predict the learning style desire of the engineering students at Atlm University [5].

The researchers have chosen the VARK model for the investigation to help educators/trainer/content creator to recognize individuals learning style to improve the learning process. VARK model unquestionably defines 4 learning styles i.e. visual, auditory, read/write and kinaesthetic. When the learning objects created by them would consist of content types Visual, Auditory, Read/Write and Kinesthetic. These attributes motivated the researchers to choose this model.

There are classifications techniques have been tried out to classify learning styles with a range of learning style models. These techniques sequence the learning activities and observe the learners behaviours with the system [20-22]. According to Dung, the courses are comprised of several topics with

different learning contents which are referred as learning objects. The learning objects are labelled based on the Felder and Silverman learning styles and not based on the percentage of the type of learning content in the Learning Object. [23]. In the proposed system classification is done dynamically based on the time spent in each type of learning content and not based on the learning sequence which is fixed by the course teacher. Its objective is to provide learner centred learning objects.

III. PROPOSED WORK

People are very much interested to know things day by day. Many platforms like web pages, videos, audio messages, illustrations and social media and so on play vital role in knowledge acquisition and sharing. e-Learning provides great opportunities to widen our knowledge through various methods. It may vary depends on user, need, task, subject and content type. The teacher, developer or content creator must take them into consideration while developing e-Learning applications and the learning objects. While developing e-Learning application or the learning content, one should clearly identify the content type. The type of learning content is one of the major attributes which should be personalized based on the learner preferred learning content type. Though there are different kinds of learning materials available in e-Learning, it is necessary to provide an appropriate content according to the learning style of the learner.

A. Architectural Design

A schematic representation is shown in Fig. 1. The system consists of two units: User Profiling and classification of learning style of the learner. The general interpretations of these units are given below.

1) *User profiling unit:* Profile is a unique trait which substantiates the success of learner in the learning activities. The profile of a learner presents the details of individual learner. The main objective of this unit is to pull together the data about the learner's behaviours in accessing the learning objects. As the learner enters to access the learning content through LMS or any educational App, the user profiling unit observes and collects the data including the number of learning objects visited, the time spent on each learning objects. The learning objects visited by the learner may be of any resource types like PPT, PDF, Demonstration, URL, Video, Audio, Images etc. Each Learning Objects are identified by its Id number and every learning object will have certain percentages of content type like Visual, Auditory Read/Write and Kinesthetic.

2) *Data pre-processing:* The time spent by the learner in each learning object is converted into percentage of time. So that learning content and the learning object are in the percentage.

3) *Classification method:* The normalized is processed for classification. The learner's pre-dominant learning preference is identified based on the time spent on each type of learning content. This facilitates the content creator to create based on the learner preferred content type.

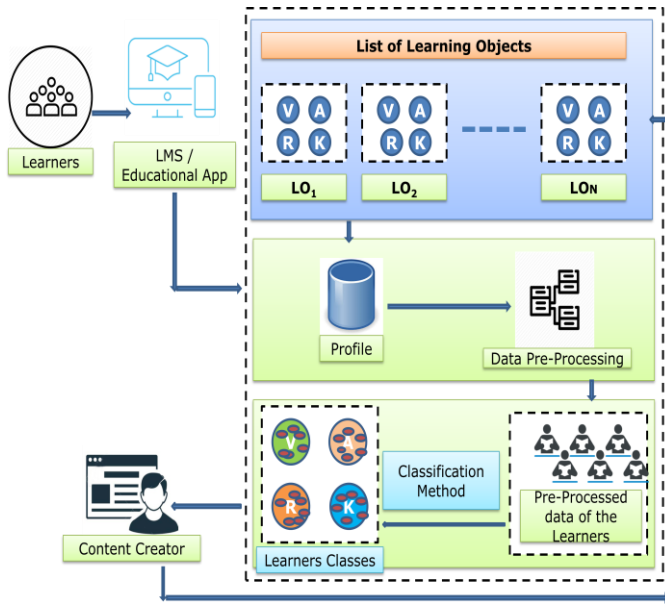


Fig. 1. Schematic Representation of Classification of Learner.

B. Methodology

The learner preferred learning style is identified as follows. As mentioned earlier the learning content is provided to the learner. The Learning Object consists of certain percentage of learning content type like Visual, Auditory, Read/Write and Kinesthetic.

$$L.O = \{lo_x \in [lo_{\%v}, lo_{\%a}, lo_{\%rw}, lo_{\%k}], x = 1, 2, 3, \dots, n\} \quad (1)$$

Where,

$L.O$ is the collection of Learning Objects

lo is the Learning Object with the vector of percentage of learning objects of type that are identified by the learning object id, $x = 1, 2, \dots, n$

$\%v$ is the percentage of Visual Content,

$\%a$ is the percentage of Auditory Content

$\%rw$ is the percentage of Read/Write Content and

$\%k$ is the percentage of Kinesthetic Content.

For a given learning object the time spent by the learner on learning content is recorded in the profile. Based on the data collected from the profile of learner, the learner's preferred learning style is identified.

$$L = \{ly \in (lo_{\%v}, lo_{\%a}, lo_{\%rw}, lo_{\%k})\}, y = (1, 2, 3, \dots, n) \quad (2)$$

Where,

L is the list of Learner who access the learning Objects

l is a learner who is identified by the user-id, $y=1, 2, 3, \dots, n$.

lo_{tv} is the time spent by the learner on visual content

lo_{ta} is the time spent by the learner on auditory

lo_{trw} is the time spent by the learner on Read/Write and

lo_{tk} is the time spent by the learner on kinaesthetic content

C. Methodology used in Classification of Learners

The steps to classify the learners are given below:

1) The learner's user-id and the learning content-id visited by the learner are observed. The following value are tabulated

a) the percentage of learning content in each Learning Object (L.O) $lo_{\%v}, lo_{\%a}, lo_{\%rw}, lo_{\%k}$ and

b) the time spent by the learner in each LO $lo_{tv}, lo_{ta}, lo_{trw}, lo_{tk}$

2) The time spent by the learner is normalized by converting it to percentage using the following equations.

$$T_V = (lo_{tv} / \sum_{i=1}^t t = lo_{tv} + lo_{ta} + lo_{trw} + lo_{tk}) * 100 \quad (3)$$

$$T_A = (lo_{ta} / \sum_{i=1}^t t = lo_{tv} + lo_{ta} + lo_{trw} + lo_{tk}) * 100 \quad (4)$$

$$T_{RW} = (lo_{trw} / \sum_{i=1}^t t = lo_{tv} + lo_{ta} + lo_{trw} + lo_{tk}) * 100 \quad (5)$$

$$T_K = (lo_{tk} / \sum_{i=1}^t t = lo_{tv} + lo_{ta} + lo_{trw} + lo_{tk}) * 100 \quad (6)$$

Where,

' t ' is the time spent by the learner in an hour (t).

T_V is the percentage of time spent by the learner on visual content

T_A is the percentage of time spent by the learner on auditory content

$T_{R/W}$ is the percentage of time spent by the learner on Read/Write content

T_K is the percentage of time spent by the learner on Kinesthetic content

3) The learner's Predominant learning style is obtained from the following equation

$$L.S_j(\bar{X}) = \text{Max} \left[\left\{ \frac{\sum lo_{tv_i}}{N}, \frac{\sum lo_{ta_i}}{N}, \frac{\sum lo_{trw_i}}{N}, \frac{\sum lo_{tk_i}}{N} \right\} \right] \quad (7)$$

Where,

$i=1, 2, 3, \dots, N$

(the number of learning objects visited by the learner)

$j = 1, 2, 3, \dots, M$ (the number of users)

The maximum value gives the predominant learning style of the learner.

4) The learning content effect factor is applied in the learning style

$$LS_f = \frac{L.S}{L.O} \quad (8)$$

Where,

L.O is the Learning Objects

L.S is the predominant Learning Style of the Learner.

$L.S_i$ is the learning style of the learner based on effect factor of the learning content.

D. Experimental Results and Discussion

The usability of the designed system is evaluated by taking 50 Learning Objects with the defined percentage of VARK content and these learning contents are identified with unique Ids. 250 undergraduate students of all gender from different departments are involved in the experiment and the amount of time spent in each learning content by the 250 learners is tabulated. A snapshot of Percentage of learning content type of Learning Object along with id numbers is shown in the table 3.

Where,

LC ID is the Learning Content Identification Number

% Visual is the percentage of Visual Content

% Auditory is the percentage of Auditory Content

% Read/Write if the percentage of Read/Write Content

% Kinesthetic is the percentage of Kinesthetic Content.

The time given for a learner is 60 minutes for a content chosen by the learner. A learner can choose Learning Object of his / her interest. The amount of time spent by the learner in each learning content for a week is tabulated. A snapshot of the dataset is shown in the Table 4.

TABLE III. A SNAPSHOT OF PERCENTAGE OF LEARNING CONTENT TYPE IN LEARNING OBJECT

LC ID	User ID	% Visual	% Auditory	% Read/Write	% Kinesthetic
16	User210	30	21	8	56
13	User232	62	23	64	99
34	User8	38	73	91	25
44	User142	35	70	48	11
50	User148	4	66	27	27
3	User17	44	35	29	79
16	User241	26	82	19	9
20	User238	11	42	10	14
17	User217	58	1	62	38
31	User203	42	47	4	15
15	User64	82	55	81	90
6	User96	40	39	46	52
20	User183	49	32	54	45

TABLE IV. A SNAPSHOT OF AMOUNT OF TIME SPENT BY LEARNING ON EACH CONTENT TYPE

LC ID	User ID	% Visual	T_V	% Auditory	T_A	% Read/ Write	T_R/W	% Kinesthetic	T-K
16	User210	30	26	21	2	8	1	56	44
13	User232	62	4	23	16	64	33	99	37
34	User8	38	51	73	53	91	25	25	4
44	User142	35	56	70	59	48	6	11	20
50	User148	4	23	66	43	27	52	27	40
3	User17	44	57	35	43	29	14	79	1
16	User241	26	39	82	37	19	1	9	36
20	User238	11	35	42	14	10	31	14	2
17	User217	58	18	1	31	62	24	38	44
31	User203	42	12	47	25	4	3	15	25
15	User64	82	33	55	60	81	12	90	6
6	User96	40	22	39	39	46	16	52	59
20	User183	49	28	32	59	54	32	45	55

TABLE V. TIME SPENT BY THE LEARNER

LC ID	User ID	% Visual	T_V	% Auditory	T_A	% Read /Write	T_R/W	% Kinesthetic	T_K	Total % LC	Total % Time
16	User1	29.86	24.44	31.25	30.00	34.38	8.89	4.51	36.67	100	100
13	User1	22.92	13.00	7.11	14.00	32.41	32.00	37.55	41.00	100	100
34	User10	34.00	42.86	8.67	11.11	26.67	34.13	30.67	11.90	100	100
44	User10	40.52	52.13	37.25	17.02	14.38	5.32	7.84	25.53	100	100
50	User10	29.96	19.13	34.41	24.35	12.55	41.74	23.08	14.78	100	100
3	User10	29.33	32.26	28.37	40.32	18.75	0.81	23.56	26.61	100	100
16	User100	58.62	20.00	8.05	10.00	8.05	20.00	25.29	50.00	100	100
20	User100	15.74	18.67	43.65	27.33	36.04	24.00	4.57	30.00	100	100
17	User100	11.76	16.26	45.10	14.63	31.37	41.46	11.76	27.64	100	100
31	User100	5.19	35.58	7.14	14.11	61.04	14.11	26.62	36.20	100	100
15	User100	33.60	41.26	18.80	8.39	14.00	34.97	33.60	15.38	100	100

TABLE VI. PREDOMINANT LEARNING STYLE

LC ID	User ID	T-V	T-A	T-R/W	T_K
16	User100	20.00	10.00	20.00	50.00
20		18.67	27.33	24.00	30.00
17		16.26	14.63	41.46	27.64
31		35.58	14.11	14.11	36.20
15		41.26	8.39	34.97	15.38
	Learner Style (LS) of User100	26.35	14.89	26.91	31.84
	% of Learning Content (LC) Chosen by User100	24.98	24.55	30.10	20.37
Total L.O (5)	Impact of L.O on L.S (L.S/L.O)	1.05	0.61	0.89	1.56

Where,

T_V is the time spent by the learner on Visual Content

T_A is the time spent by the learner on Auditory Content

$T_{R/W}$ is the time spent by the learner on Read / Write Content

T_K is the time spent by the learner on Kinesthetic Content

The learners are grouped based on the learning content visited by the learner and the amount of time spent by them is recorded and the data is normalized i.e. time is converted to the percentage. Table 5 gives the details of percentage time spent by three learners namely User1, User10 and User100.

From the table 5 the learner's predominant learning style is classified. For user100, the number of learning content visited is 5 and the time spent in percentage on each learning content is shown in the Table 6.

The result shows that the predominant learning style of User100 is Kinesthetic. This gives an idea to the teacher and

content creator to create the learner preferred learning content. The result also shows that the amount of learning content type influences the learning style.

The Fig. 2 shows the classification of learning style of first 25 users. Here we observe that in the absence of learner's predominant learning content type then the learner may be given the next predominant learning content type.

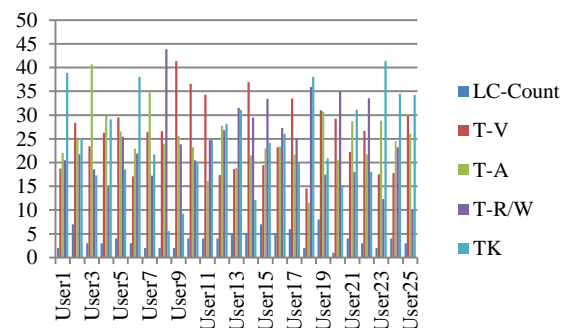


Fig. 2. Classification of Learning Style.

The classification of learners based on the impact factor of the availability of the percentage learning content type is shown in the Fig. 3.

The Fig. 4 shown below gives the clustering of individual learner i.e. number of learners belongs to Visual, Auditory, Read/Write and Kinesthetic style based on the time spent by them in the learning content type.

The Fig. 5 shown below represents the clusters of learners based on the impact factor of the learning content.

From this we observe that most of the learners of the chosen group belong to the kinaesthetic style of learners who prefer learning content in the forms of hands on exercise, case studies and demonstrations.

E. Future Work

The proposed research work and the use of VARK learning style as designed, it did not account for confounding factors such as socioeconomic status, specialization, race, culture, domicile etc. Also the number of learning content visited by the learner is highly variable, hence a weighted average on learning content count should be considered.

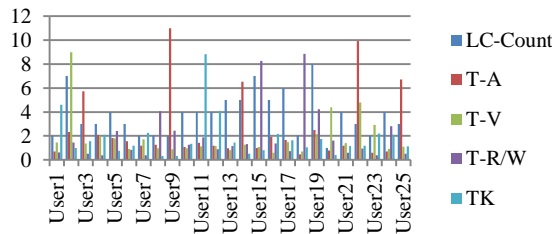


Fig. 3. Classification based on the Impact Factor of the Availability of the Percentage Learning Content Type.

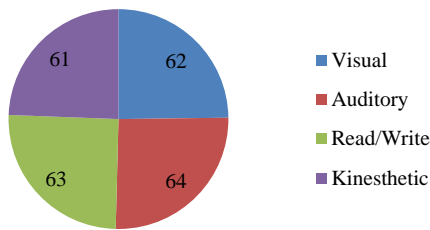


Fig. 4. Learners Cluster based on Time Spent on Learning Content.

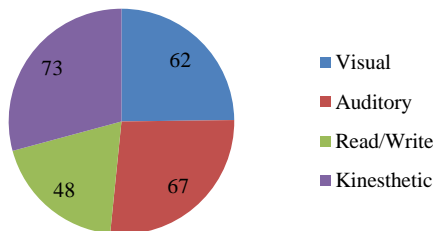


Fig. 5. Learners Cluster based on Learning Content.

IV. CONCLUSION

It is true that e-Learning environment plays a significant role in modern education. Content creators are consistently working on preparing the Learning Objects in different formats for personalized learning. One of the important features of recent LMS is to provide personalized learning object. It is a fact that the learners would have a joyful learning experience when the LMS is able to provide the learner centred learning styles and the preferred learning objects. Hence, the investigators have proposed, designed and experimented, a novel approach in the classification of learners based on the learners’ profile. When a learner enters into a new course, new semester their profile is dynamically analyzed and their learning style is classified which will facilitate the content creator to provide the learner centred learning objects. This process of classification of learner reveals that there is a high impact of learning style for the individuals to perform well in academic activities and gain higher satisfaction level.

REFERENCES

- [1] Akhras, N, F, and Self, A, J, “Modeling the Process, not the Product of Learning”, *Computer as Cognitive Tools – No More Walls*, pp 3- 28.
- [2] Beesuda Daoruang, Charun Sanrach, The Learning Material Classified Model Using VARK Learning Style, The Impact of the 4th Industrial Revolution on Engineering Education (pp.505-513), March 2020.
- [3] Bruner JS .(1967). *Towards a Theory of Instruction*, Cambridge: Harvard University Press. Retrieved from <http://www.hup.harvard.edu/catalog.php?isbn=9780674897014>.
- [4] K. Crockett, A. Latham, D. Mclean; J.O'Shea, (2013),A fuzzy model for predicting learning styles using behavioral cues in a conversational intelligent tutoring system, IEEE International Conference on Fuzzy Systems (FUZZ), 2013, pp: 1 – 8.
- [5] Burcu Devrim Ictenbas, Hande Eryilmaz, ”Determining Learning Styles of Engineering Students to Improve the Design of a Service Course”, *Procedia - Social and Behavioral Sciences (Volume 28 2011)*.
- [6] Carma Daouk, Effects of Dunn and Dunn Learning Styles Model on Achievement and Motivation: A Case Study, LEBANESE AMERICAN UNIVERSITY, School of Arts and Sciences October 2013.
- [7] Coffield, F. et al , “Learning styles and pedagogy in post-16 learning. A systematic and critical review”, *Learning and Skills Research Centre, London, 2004*.
- [8] Felder, R., and Silverman, L., (1988). Learning and teaching styles in engineering education. *Engineering Education*, 87(7), 674-684.
- [9] Harold Pashler, Mark McDaniel, Doug Rohrer, and Robert Bjork, *Learning Styles: Concepts and Evidence, Psychological Science in the PUBLIC INTEREST*, Volume 9 Number 3 December 2008.
- [10] Honey P and Mumford A (1992). *The manual of learning styles*. Maidenhead: Peter Honey Publications.
- [11] <http://www2.le.ac.uk/departments/gradschool/training/resources/teaching/theories/honey-mumford>).
- [12] Index of Learning Styles Questionnaire, <https://www.webtools.ncsu.edu/learningstyles/>
- [13] Kolb, D.A. (1984). *Experiential learning: experience as the source of learning and development* (11th ed). New Jersey-Hall.
- [14] Learning. Retrieved from [http:// en.wikipedia.org/wiki/Learning](http://en.wikipedia.org/wiki/Learning)
- [15] Norasmah Othmana , Mohd Hasril Amiruddinb , (2010). Different Perspectives of Learning Styles from VARK Model, *International Conference on Learner Diversity*.
- [16] Norman G (2009). When will learning style go out of style? *Adv Health Sci Educ Theory Pract.* 14:1–4. doi: 10.1007/s10459-009-9155-5.
- [17] R. M. Felder and J. Spurlin (2005), Applications reliability, and validity of the Index of Learning Styles. *International Journal of Engineering Education*, vol. 21, no. 1, pp. 103-112.

- [18] Robert V. Fiermonte and Kelly Bruning, (2001). Harnessing the Power of the Information Age: ELearning - New Frontier of Organizational Training, *International Journal of Instructional Technology and Distance Education*.
- [19] Zhang, L. F. (1999). Relationship between thinking styles inventory and study process questionnaire. *Personality and Individual Differences*, 29(5), 841–856.
- [20] Sheeba, T., & Krishnan, R. (2019). Automatic detection of students learning style in Learning Management System. In *Smart Technologies and Innovation for a Sustainable Future* (pp. 45–53). Springer, Cham.
- [21] Hasibuan, M. S., Nugroho, L. E., & Santosa, P. I. (2019). Model detecting learning styles with artificial neural network. *Journal of Technology and Science Education*, 9(1), 85–95.
- [22] Graf, S. (2007). *Adaptivity in Learning Management Systems Focussing on Learning Styles*. PhD thesis, Vienna University of Technology, 9801086 Neulinggasse 22/12A 1030, Vienna.
- [23] Dung, P. Q. (2012). An approach for detecting learning styles in learning management systems based on learners' behaviours. *International Conference on Education and Management Innovation*, 30.

Efficient Security Solutions for IoT Devices

Faleh Alfaleh¹ Salim Elkhediri²

Department of Information Technology
College of Computer, Qassim University, Buraydah, Saudi Arabia

Abstract—The Internet of Things (IoT) is a technological innovation that has revolutionized society. The IoT will forever change the way we use simple things that do very little things to smart, fully capable things. IoT devices can process and automate everyday household and workplace tasks through simple sensors. Yet despite the benefits of these devices, they are vulnerable to violations such as privacy issues and security breaches. This paper aims to provide a clearer understanding of the IoT and current threats to it by explaining why IoT devices are susceptible to attack. Moreover, the technologies used in the IoT are examined, as well as the different communication layers of the IoT and their functioning. The findings reveal that IoT devices are prone to many software and hardware vulnerabilities, not to mention the challenges that come with IoT. Solutions to these challenges are proposed, notably through the use of anomaly-based intrusion detection systems, which are critical components of network security. Using machine learning (ML) to detect potential attacks is recommended. Many proposed anomaly-based detection systems use different ML algorithms and techniques. However, there is no standard benchmark to compare these in terms of power consumption. A benchmark that measures both accuracy and power consumption to calculate and evaluate each algorithm's implementation is proposed.

Keywords—Efficient; IoT; Systems on a Chip (SoC); ML; Network

I. INTRODUCTION

The world is undergoing a rapid and exciting transformation because of the wide availability of systems on a chip (SoCs), as shown in Fig. 1. SoCs enable the creation of very intricate and small computer models that are able to connect to the network.

When an SoC connects to the Internet, it become the Internet of Things, forming an essential foundation for many utilities. Our everyday lives rely on their functionality and the quality of their operations. For example, industrial applications. Traditional security approaches are typically more expensive for IoT in terms of energy usage as well as overhead costs. Most security frameworks, in the event of a threat, tend to be centralized. They are therefore not appropriate for devices with a distributed network, given the difficulty of size, the existence of increased traffic and the single point of failure [2]. Acquiring data through multiple aspects of the industrial life cycle may significantly improve performance, thereby allowing a company to collect more data and monitor their industrial operations.

Moreover, many new devices can be linked to the network. Such systems tend to use most of their resources and computing power for the application's key features; thus, ensuring protection and privacy at a lower cost will be

extremely difficult. For example, the primary differentiator between elliptic curve cryptography (ECC) and Rivest Shamir Adleman (RSA) is the key size compared with cryptographic strength. ECC can deliver much smaller key sizes with the same cryptographic strength as an RSA system. A 256-bit ECC key, for example, is equal to a 3072-bit RSA key [3]. This growing threat has prompted the development of new strategies to detect and block IoT botnet attack traffic. Recent research has highlighted the promise of machine learning (ML) in identifying malicious Internet traffic [4]. Nevertheless, ML models mainly targeting IoT application networks or IoT attack flux have met with limited success. Thankfully, IoT traffic often varies from other Internet-connected products (e.g., notebooks and smartphones) [5].

The rest of this paper is organized accordingly. First, IoT layers will be presented with the wireless network technology options and the characteristics of each technology. Then the most common attacks on IoT devices and the core design challenges of IoT devices will be covered. After that, a review of recent related literature will be discussed. Then we will give an overview on the UNSW-NB15 dataset. Then, anomaly-based intrusion detection method will be introduced with six classifiers. The purposed solution will be presented with the methodology on how to evaluate Intrusion Detection System (IDS) performance. The analyzed results will be provided with and without the purposed solution. Finally, we will summarize our work and mention the future work.



Fig. 1. Raspberry Pi SoC [1].

II. INTERNET OF THINGS

Although the term “Internet of Things” is being used more frequently in everyday life, there is no universal definition of what IoT truly means. The term was first used in 1999 by Kevin Ashtonof, one of the members who created a global standard system for the Radio Frequency Identification (RFID) at the Massachusetts Institute of Technology [6][7].

A. Internet of Things Technologies

Deploying a large number of devices with limited memory and storage capabilities increases the threat to IoT applications. This is because attackers can take advantage of this weak IoT device capability to penetrate connected IoT applications. To understand the security issues related to the IoT, first we need to understand the way in which the IoT works. Some of the network technologies used in IoT:

1) *Short-Range Device (SRD)*: Short range devices or SRDs are radio frequency transmitters used to transmit information. Their ability to cause harmful interference to other radio equipment is very low. SRDs are low power transmitters; depending on the frequency range, their effective radiated power (ERP) is usually limited to 25 to 100 megawatts or less, which limits their effective range to a few hundred meters and does not require user permission. Most of the SRD protocols are considered personal area networking (PAN). Most used SDR in IoT environment:

a) *Radio Frequency Identification (RFID)*: An SRD that is frequently used in the IoT environment is RFID. This technology allows circuit boards with radio frequency design for wireless connections to transmit data. Tags perform the automatic identification of objects.

b) *Bluetooth*: Bluetooth is used for data transmission via radio waves, allowing two or more devices to connect. It is considered a short-range wireless technology. Further, Bluetooth Low Energy (BLE) protocols are well suited to the IoT because these are designed and enhanced for short-range use, low bandwidth, and low latency application. [8].

c) *ZigBee*: Zigbee has low energy consumption; therefore, it has many uses in smart homes, for example, for smart lighting. However, because its range is short, it is typically used in mesh networks where data are passed from one device to another until they reach their destination. This makes Zigbee ideal for the IoT. [9].

2) *Wireless Fidelity (Wi-Fi) IEEE 802*: Wireless Fidelity or Wi-Fi is a well-known wireless communication protocol. It offers very high data rates with a longer range than Bluetooth or RFID. IEEE 802.11ax is the most recent version [10]. with speeds reaching 600 to 9 608 megabits per second [11]. Wi-Fi can connect to various frequencies, such as 2.4, 5, 6, and 60 gigahertz. Depending on the frequencies, the range, speed, and power will vary.

3) *Cellular networks*: Cellular networks depend on the region or cell covered by the communication station. Each cell will provide the IoT application to move between sites. Example of a cellular network is the 5G communication protocol, which is an open standard under the supervision of GPP3. These networks currently support the requirements for 5G mobile communications. As of 2020, 5G networks have two types of bands which are 1- “Mid-band” uses frequencies of 2.5 to 3.7 gigahertz, currently allowing speeds of 100 to 900 megabits per second, with several miles of radius. And the “High-band” uses frequencies of 25 to 39

gigahertz to achieve download speeds of 1 to 3 gigabits per second. It only has a range of about one mile of radius.

4) *Low-Power Wide-Area Network (LPWAN)*: Low Power Wide Area Network or LPWAN is a wide-area network with low power consumption, resulting in very low speeds. This type of network was intended for use in large areas, with simple commands, such as a smart light sign. LPWAN data rates are very low, ranging between 0.3 and 50 kilobits per second with long-range communications of up to 40 kilometers. Example of LPWAN is Long Range Wide Area Network (LoRaWAN) which is a low-cost, long-range, low-power wireless platform that can be used in many IoT applications. It uses the frequencies of 433, 868, and 915 megahertz with a bit rate of 3 to 5 kilobits per second. LoRa-enabled devices can survive with a battery for years in sleep mode. In Fig. 2 we can see A comparison of Cellular networks in terms of data rate and range.

B. Internet of Things Layers

Every IoT device has at least three layers, namely, the network layer, the data processing layer, and the application layer, as shown in Fig. 3. Further, some applications have a fourth, sensing layer, not unlike a camera.

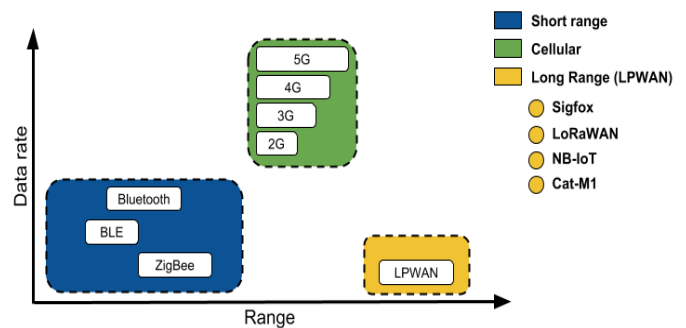


Fig. 2. A Comparison of Cellular Networks Data Rate and Range[12].

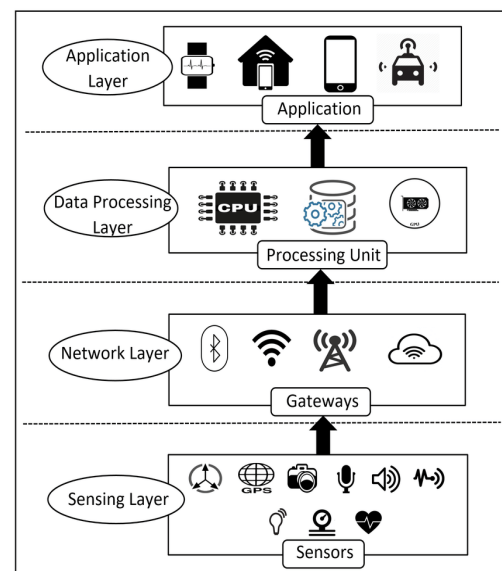


Fig. 3. IoT Architecture Layers [13].

Each layer has its unique components and role to play. These roles are not interchangeable, and each has its own technologies. The first layer contains specific applications for the IoT, for example, cloud computing platforms and middleware technology. The second layer includes data processing units such as a Central Processing Unit or Graphics Processing Unit. These are mainly used for collecting and processing data and controlling other objects. The third layer contains networks including access control, firewalls, and gates. This layer also contains technology such as 5G networks, ad hoc networks, and Wi-Fi. Different network transmissions have different technologies. Last, the sensing layer includes the technology needed to collect information, such as images, location, sound, and many other collected data from the environment. The data are then sent to the processing unit via the network layer.

C. Attacks on the Internet of Things

As the IoT evolves, the full definition of protection must be re-examined. Individuals and organizations are increasingly using IoT devices to improve productivity. The greater the number of users, the higher the chance of an attack or a vulnerability. For example, a large number of malicious nodes that used CCTV may have been part of a disseminated denial-of-service (DoS) attack from an individual home [14]. Some of the main attacks:

1) *Denial of Service (DoS) and Distributed Denial-of-Service (DDoS) Attacks:* As the IoT evolves, the full definition of protection must be re-examined. Individuals and organizations are increasingly using IoT devices to improve productivity. The greater the number of users, the higher the chance of an attack or a vulnerability. For example, a large number of malicious nodes that used CCTV may have been part of a disseminated denial-of-service (DoS) attack from an individual home [15].

2) *Spoofing attack:* Spoofing nodes impersonate the legitimate IDs of IoT devices such as media access control or RFID tags to gain illegal access to IoT systems. These attacks may launch other attacks, such as DoS and man-in-the-middle (MITM) attacks [16].

3) *Jamming attack:* In jamming attacks, the attacker sends corrupt transmitted packets to interrupt the continuous wireless transmission of the IoT device, exhausting bandwidth, power, CPU, and memory resources of the and this prevents the device from sending a signal leading to a series of system crashes. These may have serious consequences, especially in IoT applications that involve human health or internal security [17]. One example is opening a door while jamming the security sensor, as shown in Fig. 4, which results in a security breach.

4) *Man-In-The-Middle (MITM) attack:* In an MITM attack, the attacker can read and change contact between two parties believed to be communicating directly with each other. An example of an MITM attack is active eavesdropping, in which the attacker establishes independent communication with the victims and transmits messages between the victims, leading them to believe that they are talking directly to each

other via a dedicated connection. At the same time, the entire conversation are with the attacker. The attacker is thus able to intercept all relevant messages that have been passed between the two victims, and is also able to send new messages.

5) *Malware attacks:* Malware attacks consist of viruses, worms, Trojans, or rootkits. These attacks behave similarly to attacks on traditional networks. Usually, the attacker uses malware to gain sensitive data or access sensitive or critical industrial infrastructure [18].

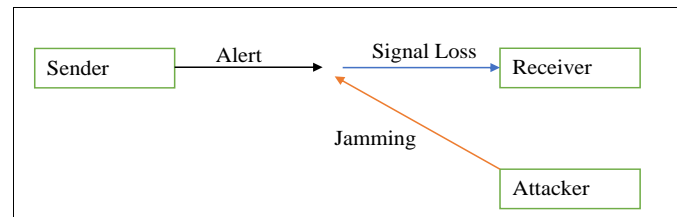


Fig. 4. Jamming Attack.

D. Challenges of The Internet of Things

The IoT is faced with many challenges from a wide variety of standards and applications, with different capabilities in terms of processing power and memory. With many traditional threats and new threats every single connection could make the networks vulnerable. Most IoT devices marketed with many features without carefully planning for security in the long term. Some of the main challenges of the IoT are described in the following sections.

1) *Lack of standard:* There are no standard IoT each device with its unique spaces and technology; most IoT devices that are released onto the market have at least one different wireless module or type of controller.

2) *Privacy concerns:* Recent developments in the IoT have introduced IoT devices into our homes, our doors, our cars—even into stores, such as Amazon’s self-service grocery stores [19]. Therefore, it is becoming increasingly difficult to protect personal privacy and prevent the unsolicited collection of personal information. Moreover, different attacks can violate personal identity and location [20].

3) *Distributed nature:* The ability to distribute devices according to need and required distance means it is difficult to manage a large number of distributed devices.

4) *Insecure physical interface:* Several physical factors compound the threats to the proper functioning of IoT devices. This is especially so when sensors and equipment are installed in a public area, making them vulnerable to physical attack.

5) *Scalability:* The IoT is a scalable technology, which creates many challenges, for example, the need for scalable technology and algorithms.

6) *Specification:* Each IoT device has different capabilities in terms of processing power, storage, RAM, and battery. For that, considering what type of security solution used is critical.

7) *Real-Time:* IoT devices are required to be used in real time. Because sensors are included in IoT systems, fast and stable sensor is a must to ensure continuous real-time

performance. Therefore, even for embedded devices with limited functionality, the IoT system must support the device or user in real time.

III. LITERATURE REVIEW

A literature review of recent works on the security in resource-constrained devices like IoT devices. These devices remain one of the most cost-effective solutions for many day to day applications. The quantity of devices connected to the Internet continues to grow at a steady pace. A recent forecast from the International Data Corporation estimates that there will be 41.6 billion connected IoT devices, generating 79.4 zettabytes of data in 2025 [21].

A. Related Work

Sicari et al.[22] The authors discuss the confidentiality, authentication, data security issues, network security, and intrusion detection systems and the continuing lack of communication standards. Proper implementations must be developed and implemented regardless of the system used to guarantee security, access control, and the privacy of users and objects as well as the performance of the devices Compliance of specific policies on security and data protection. Despite many attempts in this field, many challenges and research problems remain. In particular, the author maintains that there is still a lack of systems and a unified vision to ensure the security of the Internet of Things. Then the author provides an analysis of international projects in this field, indicating that these efforts usually aim to design and implement specific applications of the Internet of Things. The study is also concerned with the need to address the use of IoT technologies, also communications into protected middleware, capable of meeting specific security constraints.

Hongchun et al.[23] proposed a knowledge-based intrusion detection strategy to detect multiple types of attacks under different types of network structures. The purpose was to create an independent detection model that depends on the structure of the WSN network. The suggested mechanism was based on the fact that different types of attacks may have different forms of density. The authors collected network traffic and used it as a feature of random network behavior in the feature space. The density form can be considered an indication of normal and abnormal network behavior. Simulation results from attacks, such as a sinkholes, flooding,

or DoS, indicate that the method had the appropriate detection accuracy and high compatibility with the network structure.

Stergiou et al.[24] Present the IoT with a cloud computing survey that reflects on and how to secure the security problems on both technologies they specifically combine the two technologies as mentioned earlier (i.e., cloud and computing and IoT) to examine the usual attributes and to find out about the advantages of their integration. They demonstrate how the security problems of IoT integration can be strengthened by cloud computing. The theoretical application design and the integration of IoT and Cloud Computing with security benefits are further analyzed by the two encryption algorithms which are used (AES and RSA).

Doshi et al.[25] Presented that high precision DDoS attacks in IoT traffic may be identified with several machine learning algorithms, including neural networks, through the use of IoT network behavior to notify feature attacks. The results suggest that main gateways or other central network boxes will classify locally based IoT DDoS attack sources automatically using inexpensive machine learning algorithms and an independent flow-based traffic-based data protocol. DoS identification can reliably differentiate between usual and DoS attack traffic by using the packet level machine learning for IoT consumer devices. They used a small set of features to reduce computational overhead, which is essential for the real-time identification and deployment of the middlebox. Their selection of features was based on the assumption that network traffic habits for IoT applications clients differentiate from those of well-known non-IoT networked devices. The test array accuracy of all five algorithms reached 0.99. The way they test this is shown in Fig. 5. These initial findings inspire further studies on anomaly machine learning to protect IoT devices.

Damopoulos et al. [26] discusses the importance of IDS for mobile devices and the importance of personal files designed for each user in order to create an effective IDS to prevent an attack. They measured several algorithms in the Phone activity data set they built and recorded the results. The authors found that they could detect anomaly with high accuracy. They also collected some useful indicators for each algorithm used in mobile phone identifiers. Their main focus was on creating IDSs that can be used with the data set and specifically for an anomaly.

Moreover, a related work comparison for all previous studies was provided in Table I.

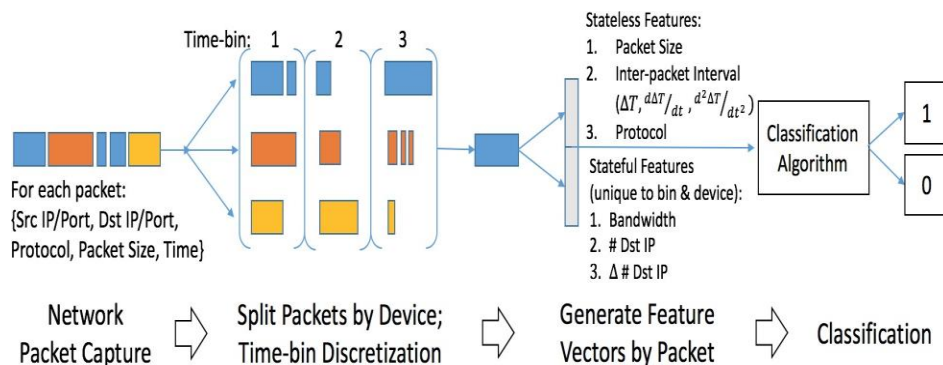


Fig. 5. IoT DDoS Detection Pipeline [25].

TABLE I. RELATED WORK COMPARISON

Paper Title	About	Advantage	Disadvantage
Security, privacy and trust in the Internet of things: The road ahead [22]	Internet of Things Survey	Present challenges and the existing solutions that may help in the field of IoT security.	The authors do not indicate in deep the physical challenges faced IoT in terms of resources.
An Adaptive Intrusion Detection Method for Wireless Sensor Networks [23]	IDS	The authors proposed a knowledge-based intrusion detection strategy (KBIDS) to detect multiple forms of attacks.	The authors proposed IDS for the WSN generally, but they do not consider the limited resources.
Secure integration of IoT and Cloud Computing [24]	Cloud Cryptography	The authors suggested cloud computing as a solution for IoT integration to processing and dealing with data.	The authors did not discuss their solution practically.
Machine learning DDoS detection for the consumer internet of things devices [25]	IDS	The authors purposed an IDS that detect DDoS attack by using lightweight machine learning algorithms.	The authors only discuss the DDoS attack.
Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers [26]	IDS	The author reviewed and present the important of using IDS on mobile devices and how it impacts on discovering anomaly.	The author only suggests a standard that can be used for future research.

IV. UNSW-NB15 DATASET

The UNSW-NB15 dataset was designed at the Cyber Range Lab of the Australian Centre for Cyber Security at the University of New South Wales [27].

A. Why UNSW-NB15

UNSW-NB15 was chosen because it is one of the most recent datasets, compared to using older datasets such as the KDD Cup 99 dataset and the NSLKDD dataset, which lack new low-fingerprint attack methods and do not include the most recent normal traffic scenarios. As a result, it can accurately represent both traditional network traffic and multiple botnets cyberattacks. IXIA PerfectStorm was used to create the dataset. This tool mixes legitimate user network traffic with malicious network traffic [27].

B. UNSW-NB15 Attacks

In this dataset, there are nine types of attacks, namely, Fuzzers, Analysis, Backdoor, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms, as shown in Table II.

The detailed number of instances in each category can be found in Table III.

TABLE II. ATTACK TYPES [27]

Attack	Description
Exploit	This attack exploit a glitch, bug, or vulnerability of a host or network.
Fuzzers	This is an attack that tries to discover security loopholes in a system and by flood it with random data until it crashes.
DoS	This attack disrupts the computer resources via flood the system with requests, making it too busy to be accessing a device.
Analysis	This is a type of intrusion that attacks web applications via ports, emails, or web scripts.
Backdoor	This is a technique of stealthily bypassing authentication, and provide unauthorized remote access.
Reconnaissance	This can be defined as a probe; probing attacks involve a method to gain information about a network, for example, port scanning.
Generic	This is a technique used against block cipher using a hash function to collision without looking how the configuration of the block cipher.
Shellcode	This is an attack in which the attacker exploit vulnerability in a program to open remote shell to control the compromised machine.
Worm	This is an attack that can replicates itself to spread to other computers via the network.

TABLE III. NUMBER OF INSTANCES OF EACH ATTACK TYPE OF UNSW-NB15 DATASET [27]

Category	Total number
Normal	93 000
Analysis	877
Backdoor	2 329
DoS	16 353
Exploits	44 525
Fuzzers	24 346
Generic	58 871
Reconnaissance	13 987
Shellcode	1 511
Worms	174
Total number of attacks	164 673
Total	257 673

V. ANOMALY USING MACHINE LEARNING

With the growing popularity of the Internet and the widespread use of computers and IoT devices, the opportunities for attacks have increased in smart and industrial IoT applications. It is difficult to counter this problematic environmental advantage using traditional techniques to detect traffic anomalies. This emerging threat has prompted the development of new techniques to identify and block attacks. In this paper, supervised learning techniques were used, namely, the Decision Table, K-nearest neighbor (K-NN), Decision Tree, LogitBoost, Naive Bayes and Random Forest.

A. Machine Learning and Classification Algorithms

The ML technique known as classification is used to distinguish attacks or intrusions from ordinary events that occur in the network. It analyzes a given dataset and assigns the instance to a particular class to minimize classification error and extract models that accurately define important data classes within the given dataset. Most ML algorithms can be classified according to the expected structure of the model. In this paper, the focus is on classification. This refers to ML algorithms that are provided with a labeled training dataset. In this paper, the UNSW-NB15 training dataset was used to build the classification model. Fig. 6 shows two types of ML. The first uses supervised ML and the second uses unsupervised ML.

For classification, six classifiers were chosen because of their accuracy while maintaining a reasonable time for testing. These are discussed below.

1) *Decision tree*: A decision tree resembles a flowchart; it uses a supervised classification technique and consequently, it requires a labeled training dataset such as the UNSW-NB15 dataset to construct a decision tree. This is done by repeating the input data through a learning tree [29].

2) *Random forest*: A random forest is a mixture of tree predictors. Each tree depends on random vector values that are sampled independently, with the same distribution for all trees in the forest [30].

3) *Decision table*: A decision table is a descriptive visual representation of actions to be performed based on conditions. An algorithm's output represents a set of operations that build in the decision-making table. A decision table can be used to describe and analyze a situation. The decision is taken based on the number and interrelationships of conditions [31].

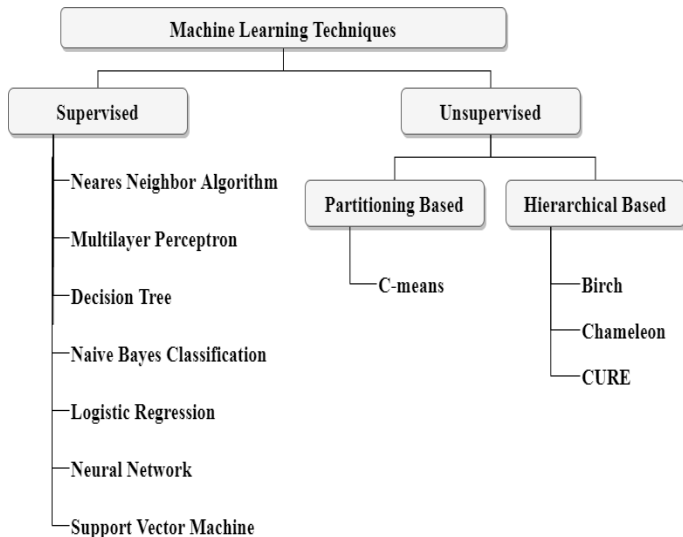


Fig. 6. Machine Learning Techniques[28].

4) *K-nearest neighbour (K-NN)*: The K-NN classifier uses a supervised learning algorithm. This algorithm does not build a model; instead, it uses a distance measure to locate K. "Close" instances in the training data for each test instance

and uses those selected instances to make a prediction. This function is calculated concerning K the nearest point, so the K-nearest neighbors do not need high computing power to run. This factor, in addition to the relative readings of adjacent nodes, makes neighboring neighbors a distributed learning algorithm suitable for WSNs [32].

5) *Naive bayes*: Naive Bayes is a simple probability classifier used to represent binary and multi-class classification problems. It is assumed that the value of a variable affects a specific class independently of the values of other variables. This assumption reduces the number of parameters. In practical terms, this is not a serious problem because even if this assumption does not apply to the data being analyzed, a naive Bayesian model performs well while significantly reducing computation time without sacrificing performance [33].

6) *LogitBoost*: This is an algorithm from the "boosting" category. It works by training a series of weak models (e.g., regression, boosting algorithms focus on increasing the ability of prediction). This algorithm was written by Jerome Friedman, Trevor Hastie, and Robert Tibshirani in 1998 [34]. It was designed to address the ability of AdaBoost to deal with noise and outliers. LogitBoost's algorithm uses a probability binomial logarithmic equation, which changes the loss function linearly. In comparison, AdaBoost uses the exponential loss function, which changes greatly with classification error. Therefore, LogitBoost is more effective to outliers and noise in general.

VI. EVALUATION METHODOLOGY

The methodology on how to evaluate IDS ML that operates in the IoT environment, comparing the accuracy results while also taking into account efficiency. To use ML, compare different algorithms and evaluate the performance of each algorithm in the IDS, the evaluation methods discussed below were used.

A. Performance Measurement

IDS creates alarms (intrusion) detecting general conditions of attack and normal behaviour. This behaviour is identified as

- *True-positive (TP)*: The number of actual attacks detected.
- *True-negative (TN)*: The number of regular activities detected as normal.
- *False-positive (FP)*: (intrusion Missed) The number of attacks detected as regular traffic.
- *False-negative (FN)*: The number of regular activities detected as an attack.

Table IV shows a simple metrics to identify each of the classifier output.

The percentage number of true alarms and false alarms for each classifier is then provided and the correct number of "Intrusions Detected" or "Intrusions Missed" measured. Then the overall accuracy of the classifier is indicated.

$$\text{Accuracy} = \frac{(TP + TN)}{\text{total number of instances}} \quad (1)$$

$$\text{Missed Intrusions} = \frac{\text{missed intrusions}}{\text{total number of intrusions} * 100} \quad (2)$$

B. Feature Selection

Selecting the features from a dataset is a way of improving the efficiency of ML algorithms. Some of the data in the dataset are irrelevant, redundant, or noisy features. Feature selection reduces the number of features by removing irrelevant, redundant, or noisy features. Feature selection speeds up the ML algorithm; it can improve learning accuracy, and lead to better model comprehensibility [35]. As shown in Fig. 7 the info gain attribute evaluation method was used, which is a Filter feature selection method that uses statistical techniques to evaluate the relationship between each input variable and the target variable with low computation power. It employs a ranked system that assigns a value from 0 to 1 for each attribute. Those attributes with higher value have a higher information value and can be selected in the optimal feature, while those with lower information value are removed.

C. Evaluation of Power Consumption

Metrics regarding power consumption were collected, as in the time, the classifier needs first to build the model then testing it. After that, the time needed to finish the task will be used to calculate the power consumption.

TABLE IV. CONFUSION METRICS

	Is the IDS correct?	
	Yes	No
Attack	True positive (TP) = Actual Attacks	False-positive (FP) = Intrusion Missed
Normal	True negative (TN) = Actual Normal	False-negative (FN) = False Alarm

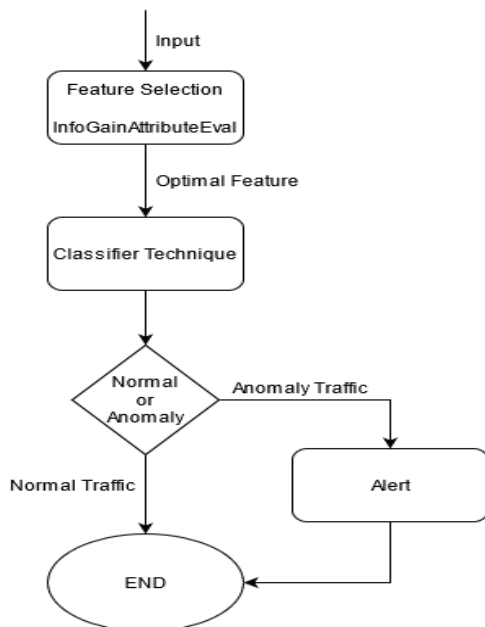


Fig. 7. Feature Selection Techniques Model for Classification.

Building Time: Time taken in the learning phase when the model is constructed from the network traffic dataset.

Testing Time: Time taken in the detection phase, showing the efficiency of the IDS.

VII. IDS PERFORMANCE RESULT

As discussed previously, the evaluation method on the UNSW-NB15 dataset was applied on the Decision Table, K-NN, Decision Tree, LogitBoost, Naive Bayes, and Random Forest.

A. Evaluation Metrics

For comparison, metrics were evaluated, as discussed. Below tables show the result parameters taken from the testing phase for the Decision Table, K-NN, Decision Tree, LogitBoost, Naive Bayes and Random Forest approaches respectively. Table V to Table VIII shows the evaluation metrics result of the original dataset and after applying Feature Selection.

TABLE V. EVALUATION METRICS PERCENTAGE OF THE ORIGINAL DATASET

Classifier	TP Rate	FP Rate	TN Rate	FN Rate
Decision Table	98.50%	5.20%	94.80%	1.50%
K-NN	96.10%	5.40%	94.60%	3.90%
Decision Tree	98.70%	1.60%	98.40%	1.30%
LogitBoost	94.00%	12.60%	87.40%	6.00%
Naive Bayes	67.20%	9.80%	90.20%	32.80%
Random Forest	98.80%	2.40%	97.60%	1.20%

TABLE VI. EVALUATION METRICS PERCENTAGE AFTER APPLYING FEATURE SELECTION

Classifier	TP Rate	FP Rate	TN Rate	FN Rate
Decision Table FS	97.80%	5.10%	94.90%	2.20%
K-NN FS	97.80%	2.40%	97.60%	2.20%
Decision Tree FS	98.50%	1.50%	98.50%	1.50%
LogitBoost FS	97.50%	24.20%	75.80%	2.50%
Naive Bayes FS	87.10%	26.50%	73.50%	12.90%
Random Forest FS	98.50%	2.00%	98.00%	1.50%

TABLE VII. EVALUATION METRICS OF THE ORIGINAL DATASET

Classifier	Actual Attacks	Intrusion Missed	False Alarm	Actual Normal
Decision Table	40723	623	1194	21878
K-NN	39715	1631	1244	21828
Decision Tree	40806	540	379	22693
LogitBoost	38871	2475	2900	20172
Naive Bayes	27783	13563	2270	20802
Random Forest	40854	492	554	22518

TABLE VIII. EVALUATION METRICS AFTER APPLYING FEATURE SELECTION

Classifier	Actual Attacks	Intrusion Missed	False Alarm	Actual Normal
Decision Table FS	40450	896	1174	21898
K-NN FS	40438	908	551	22521
Decision Tree FS	40714	632	342	22730
LogitBoost FS	40311	1035	5591	17481
Naïve-Bayes FS	35995	5351	6119	15953
Random Forest FS	40731	615	465	22607

B. Missed Intrusions

From the data collected from the test, the missed intrusions were calculated. The difference in the percentage of missed intrusions was between the original dataset and after applying feature selection.

As shown in Fig. 8, all six classifier algorithms were tested. In terms of missed intrusions, the detailed analytical results show the number of missed intrusions and the percentage of missed intrusions compared with the total number of intrusions. The feature selection was conducted with the Info Gain Attribute Eval algorithm. The result for the Naive Bayes and LogitBoost classifiers comes with a massive improvement in terms of intrusion detection. Testing Naive Bayes on the original dataset caused 13 563 (32.80%) intrusions to be missed while LogitBoost caused 2 475 (5.99%) intrusions to be missed. The result after applying the feature selection on the dataset improved the Naive Bayes’ ability to detect intrusions to 5 351 (12.94%) and LogitBoost to 1 053 (2.50%). For the other four classifiers, the difference between the original dataset and feature selection the change in missed detection is minimum.

C. Accuracy

In terms of accuracy, the best detection was achieved by obtaining accuracy as close to 100% as possible. The feature selection was applied to the dataset using Info Gain Attribute Eval via the ranked selection method. The 13 most useful features were chosen.

As shown in Fig. 9, there were six classifiers algorithms. The detailed analytical results were for the accuracy of each classifier with the two parameters of performance, namely, accuracy and accuracy FS (Feature Selection). FS means that the feature selection was applied on the dataset. As a result of the six parameters with the Info Gain Attribute Eval algorithm, the FS method shows that the Decision Tree had the highest accuracy of all six classifiers at 98.57%. This accuracy decreased, but after applying the FS, it still had the highest accuracy at 98.49%. Thereafter, Random Forest was used at an accuracy of 98.37% and 98.32% after FS was applied, resulting in only a 0.05% loss of accuracy. The same results were recorded for the Decision Table with an accuracy of 97.17% and 96.78%. For K-NN, the result had increased accuracy after applying the FS, rising from 95.53% to 97.73%, which is a 2.20% improvement. Then LogitBoost was tested, with decreased accuracy recorded after FS was applied at a reduction of 1.94%, which dropped from 91.65% to 89.71%. Finally, Naive Bayes was tested, which showed the greatest improvement after FS was applied at 6.77%. This carried the result from 75.42% to 82.19%.

D. Power Consumption

In this test, the build time is calculated, which is the time it takes for the ML algorithm to build a model in the building phase, and also, the test time, which is the time it takes in the detection phase, as shown in Table IX.

Fig. 10 shows the time taken to build a model.

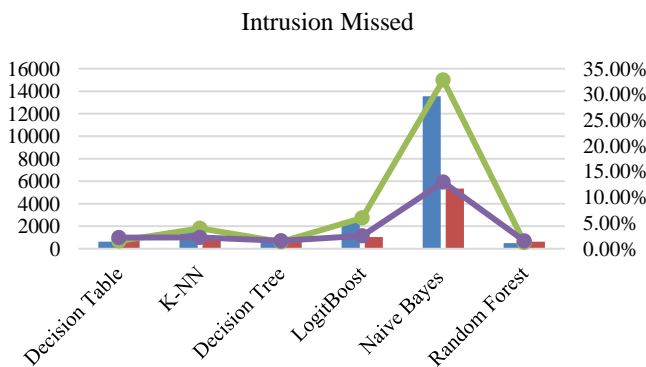


Fig. 8. Intrusion Missed.

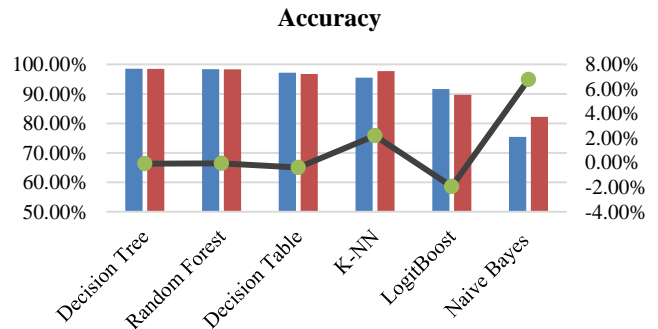


Fig. 9. Classifier Accuracy.

TABLE IX. CLASSIFIER BUILD \ TEST TIME

Classifier	Build Time	Build Time (FS)	Test Time	Test Time (FS)	Total Time	Total Time (FS)
Decision Table	391.7	162.56	0.24	0.45	392.15	163
K-NN	0.17	0.04	1529	1701	1701.8	1701
Decision Tree	73.66	23.83	0.27	0.06	73.72	23.89
Logit-Boost	37.15	11.82	0.37	0.31	37.46	12.13
Naive Bayes	2.95	0.83	1.81	0.47	3.42	1.3
Random Forest	315.	201.72	3.85	2.94	318.6	204.6

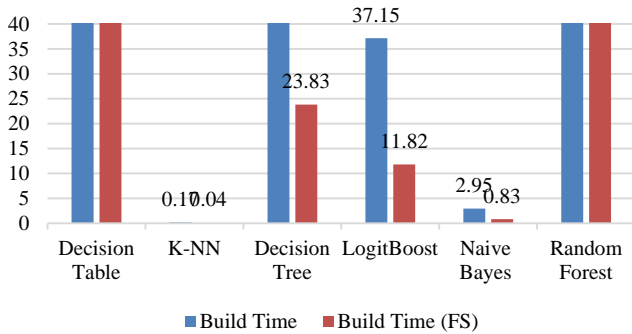


Fig. 10. Build Time (Seconds).

After building the model, the classifier used the model to test the accuracy of detecting attacks. The result shown in Fig. 11 shows the time to complete a test. The test phase considered more important because it shows the time needed from the device to finish the test and correspond directly to power consumption. As shown the decision tree has the fastest test time after applying FS at 0.06s from 0.27s, same result goes on to logitboost as it achieved 0.31s after applying FS from 0.37s. The Naïve Bayes has a mass boost in speed after applying FS as it was 1.81s to 0.47s, the result continues with random forest as it gain preforms boost after applying FS from 3.85s to 2.94s. Some classifiers loss some preforms after applying FS, K-NN decreased preforms as it was 1529.38s and after applying the FS the time increased to 1701.7, same result goes to Decision Table as it was 0.24s to 0.45s after applying FS.

E. Result Conclusion

Fig. 12 shows the fastest three tests of six. The results indicate the impact of using FS in terms of time and accuracy changes.

As can be seen in Fig. 12, using FS in terms of time has improved considerably. As discussed previously, the use of FS helps to remove any unhelpful data to improve the power consumption. The time needed to test the model in the Decision Tree classifier decreased from 0.27 to just 0.06 seconds, which was the fastest classifier with the highest accuracy of all the six classifiers that were tested. There was minimal reduction in terms of accuracy because of the FS technique.

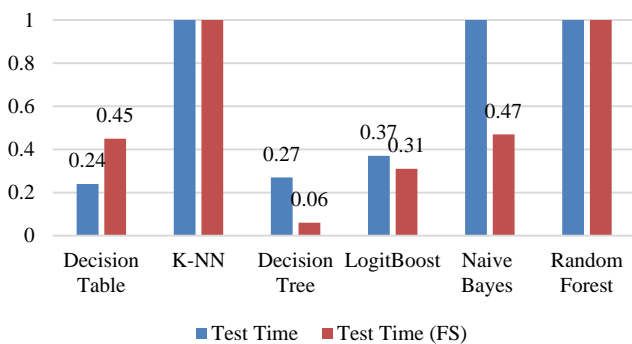


Fig. 11. Test Time (Seconds).

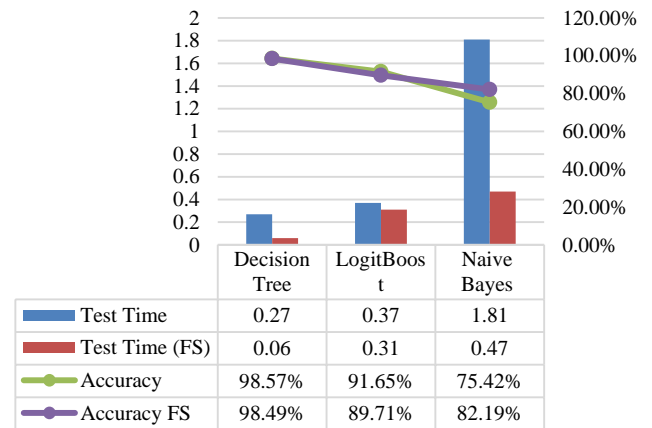


Fig. 12. Test Time (Second) with the Accuracy for the Selected Classifiers.

Moreover, in IoT applications, real-time networking, and power consumption is key. This means the important result is in the test time, which is the time it takes the IDS to test incoming traffic. The Decision Tree had the lowest time in terms of testing the incoming traffic, at only 0.6 seconds, after applying the FS. This resulted in much lower total CPU usage and battery consumption while maintaining the highest accuracy.

VIII. CONCLUSION

The aim of this paper was to provide an overview of several algorithms, implemented in a constrained environment, while maintaining protection for the IoT environment. The paper demonstrated how supervised ML could be applied to analyze network traffic data to detect intrusion accurately. It demonstrated the efficiency of the method in terms of selecting the important features to speed up training and testing time. Specific use cases focus on metrics. In contrast, the aim was to identify the most efficient classifier. This test provides definitive numbers that can be used to compare these algorithms. The results demonstrated the advantages and disadvantages of each algorithm used for anomaly-based IDS.

IX. FUTURE WORK

With the development of the Internet of Things with many distinctive features, it has put the IoT in a situation where standards and specifications for these devices are very different from any traditional solutions. For that, the available traditional solutions are not suitable for the IoT environment. Furthermore, the architecture of IoT environment usually made with arm environment that are way different than traditional x86. Moreover, the rapid growth of IoT with unique specifications has placed us in a situation where more research on efficient security solutions that suit most IoT is a must.

REFERENCES

- [1] "System on a chip - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/System_on_a_chip. [Accessed: 12-Mar-2020].
- [2] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," *Comput. Networks*, vol. 57, no. 10, pp. 2266–2279, Jul. 2013, doi: 10.1016/j.comnet.2012.12.018.

- [3] F. Alfaleh, H. Alfehaid, M. Alanzy, and S. Elkhediri, "Wireless Sensor Networks Security: Case study," 2019, pp. 1–4, doi: 10.1109/cais.2019.8769510.
- [4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3. 01-Jul-2009, doi: 10.1145/1541880.1541882.
- [5] N. Apthorpe, D. Reisman, and N. Feamster, "A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic," May 2017.
- [6] F. Wortmann and K. Flü, "Internet of Things Technology and Value Added," *Bus. Inf. Syst. Eng.*, doi: 10.1007/s12599-015-0383-3.
- [7] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Comput. Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010, doi: 10.1016/j.comnet.2010.05.010.
- [8] F. Samie, L. Bauer, and J. Henkel, "IoT Technologies for Embedded Computing: A Survey," doi: 10.1145/2968456.2974004.
- [9] S. C. Ergen, "ZigBee/IEEE 802.15.4 Summary," 2004.
- [10] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A tutorial on IEEE 802.11ax high efficiency WLANs," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 1, pp. 197–216, Jan. 2019, doi: 10.1109/COMST.2018.2871099.
- [11] "Wi-Fi - Wikipedia." [Online]. Available: <https://en.wikipedia.org/wiki/Wi-Fi>. [Accessed: 14-Mar-2020].
- [12] "Connectivity Now and Beyond; exploring Cat-M1, NB-IoT, and LPWAN Connections." [Online]. Available: <https://ubidots.com/blog/exploring-cat-m1-nb-iot-lpwan-connections/>. [Accessed: 24-May-2020].
- [13] A. K. Sikder, G. Petracca, H. Aksu, T. Jaeger, and A. S. Uluagac, "A Survey on Sensor-based Threats to Internet-of-Things (IoT) Devices and Applications," Feb. 2018.
- [14] "Threat Advisory: Mirai Botnet | Akamai." [Online]. Available: <https://www.akamai.com/us/en/resources/our-thinking/threat-advisories/akamai-mirai-botnet-threat-advisory.jsp>. [Accessed: 11-Nov-2019].
- [15] J. Fruhlinger, "The Mirai botnet explained: How IoT devices almost brought down the internet," *CSO Online*, Mar. 2018.
- [16] L. Xiao, Y. Li, G. Han, G. Liu, and W. Zhuang, "PHY-Layer Spoofing Detection with Reinforcement Learning in Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10037–10047, Dec. 2016, doi: 10.1109/TVT.2016.2524258.
- [17] R. Halloush, "Transmission Early-stopping Scheme for Anti-jamming over Delay-sensitive IoT Applications (IEEE Internet of Things Journal) Transmission Early-stopping Scheme for Anti-jamming over Delay-sensitive IoT Applications," 2019, doi: 10.1109/JIOT.2019.2911683.
- [18] S. Sharmeen, S. Huda, J. H. Abawajy, W. N. Ismail, and M. M. Hassan, "Malware Threats and Detection for Industrial Mobile-IoT Networks," *IEEE Access*, vol. 6, pp. 15941–15957, Mar. 2018, doi: 10.1109/ACCESS.2018.2815660.
- [19] "Amazon opens a supermarket with no checkouts - BBC News." [Online]. Available: <https://www.bbc.com/news/business-42769096>. [Accessed: 14-Mar-2020].
- [20] J. Zhou, Z. Cao, X. Dong, and A. V. Vasilakos, "Security and Privacy for Cloud-Based IoT: Challenges," *IEEE Commun. Mag.*, vol. 55, no. 1, pp. 26–33, Jan. 2017, doi: 10.1109/MCOM.2017.1600363CM.
- [21] "The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast." [Online]. Available: <https://www.idc.com/getdoc.jsp?containerId=prUS45213219>. [Accessed: 29-Mar-2020].
- [22] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of things: The road ahead," *Computer Networks*, vol. 76. Elsevier B.V., pp. 146–164, 15-Jan-2015, doi: 10.1016/j.comnet.2014.11.008.
- [23] H. Qu, Z. Qiu, X. Tang, M. Xiang, and P. Wang, "An Adaptive Intrusion Detection Method for Wireless Sensor Networks," 2017.
- [24] C. Stergiou, K. E. Psannis, B. G. Kim, and B. Gupta, "Secure integration of IoT and Cloud Computing," *Futur. Gener. Comput. Syst.*, vol. 78, pp. 964–975, Jan. 2018, doi: 10.1016/j.future.2016.11.031.
- [25] R. Doshi, N. Apthorpe, and N. Feamster, "Machine learning DDoS detection for consumer internet of things devices," in *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, 2018, pp. 29–35, doi: 10.1109/SPW.2018.00013.
- [26] D. Damopoulos, S. A. Menesidou, G. Kambourakis, M. Papadaki, N. Clarke, and S. Gritzalis, "Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers," *Secur. Commun. Networks*, vol. 5, no. 1, pp. 3–14, Jan. 2012, doi: 10.1002/sec.341.
- [27] J. Slay and N. Moustafa, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Artic. Inf. Secur. J. A Glob. Perspect.*, 2016, doi: 10.1080/19393555.2015.1125974?tab=permissions.
- [28] N. A. Mahadi, M. A. Mohamed, A. I. Mohamad, M. Makhtar, M. F. A. Kadir, and M. Mamat, "A Survey of Machine Learning Techniques for Behavioral-Based Biometric User Authentication," in *Recent Advances in Cryptography and Network Security, InTech*, 2018.
- [29] C. Modi, D. Patel, B. Borisanya, A. Patel, and M. Rajarajan, "A novel framework for intrusion detection in cloud," in *Proceedings of the 5th International Conference on Security of Information and Networks, SIN'12*, 2012, pp. 67–74, doi: 10.1145/2388576.2388585.
- [30] B. Kumar Baradwaj, R. Scholor, S. Pal, and S. Lecturer, "Mining Educational Data to Analyze Students' Performance," 2011.
- [31] M. Kryszkiewicz, "Rough set approach to incomplete information systems," 1998.
- [32] P. P. Jayaraman, A. Zaslavsky, and J. Delsing, "Intelligent processing of K-nearest neighbors queries using mobile data collectors in a location aware 3D wireless sensor network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6098 LNAI, no. PART 3, pp. 260–270, doi: 10.1007/978-3-642-13033-5_27.
- [33] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 2, pp. 1153–1176, Apr. 2016, doi: 10.1109/COMST.2015.2494502.
- [34] J. Friedman, J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Ann. Stat.*, vol. 28, p. 2000, 1998.
- [35] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "The Fourth Workshop on Feature Selection in Data Mining."

Internet of Things (IoT) based Smart Vehicle Security and Safety System

Yassine SABRI¹

Laboratory of Innovation in Management
and Engineering for Enterprise (LIMIE),
ISGA Rabat, 27 Avenue Oqba,
Agdal, Rabat, Morocco

Aouad Siham²

Mohammed V University of Rabat
Smart Systems Laboratory (SSL)
ENSIAS, Morocco

Aberrahim Maizate³

RITM- ESTC/CED -ENSEM,
University Hassan II Km7,
El jadida Street, B.P.
8012, Oasis, Casablanca 8118

Abstract—The Internet of Things (IoT) is making human life easy in all aspects. The applications it offers are beyond comprehension. IoT is an abstract idea, a notion which interconnects all devices, tools, and gadgets over the Internet to enable these devices to communicate with one another. IoT finds application in various areas, such as intelligent cars and their safety, security, navigation, and efficient fuel consumption. This project puts forth a solution to achieve the desired outcome of saving precious human lives that are lost to road crashes. In this context, we propose to develop a system, we are designing and deploying a system that not only avoids accidents but also to take action accordingly. This research aims at dealing with the issues that cause fatal crashes and also integrates measures to ensure safety. Life without transportation is impossible to imagine; it makes far off places easy to reach and greatly reduces the travel time. But the problems which surface due to the ever-increasing number of vehicles on the road cannot be ignored. The project aims to eradicate a few of the major reasons of car crashes and also aims to integrate post-crash measures.

Keywords—Smart vehicle security; safety system; Internet of Things (IoT)

I. INTRODUCTION

Before the discovery of the wheel, primitive man would remain secluded from other groups and communities. They could commute only within walking distance. The discovery of the wheel entirely evolved the early man life. His social boundary also grew with time. With passing time, primitive man evolved to a mannered, civilized individual and refined the design of the wheel. With the advent of technology, transportation has become an indispensable part of our lives. Though it has countless advantages and uses, we have to deal with the major problem it brings with it that costs human life. Statistically, according to Ministry of Statistics and Programme Implementation, there were 114 million motor vehicles registered in India in the year 2009 and 159 million in the year 2012. The data provided by Delhi Statistical Hand Book clearly indicates the rise in the number of registered motor vehicles from 534,000 to 877,000 in the year 2014–2016 thus increasing the number of accidents and in turn the casualties associated with the surge. Data collected by the National Crime Bureau and Ministry of Road Transport and Highway revealed that in the year 2013 more than 100,000 people lost their lives in road rage. Despite the efforts of awareness campaigns, road signs, and traffic rules, motor accidents accounted for 83% of total traffic-related bereavements in the year 2015 as published by IndiaSpand.

A. Motivation

The Internet of Things (IoT) is making human life easy in all aspects. The applications it offers are beyond comprehension. IoT is an abstract idea, a notion which interconnects all devices, tools, and gadgets over the Internet to enable these devices to communicate with one another. It utilizes information technology, network technology, and embedded technology. Various sensors and tracking devices are coupled to deliver the desired outcome thus making lives easier. IoT finds application in various areas, such as intelligent cars and their safety, security, navigation, and efficient fuel consumption. This project puts forth a solution to achieve the desired outcome of saving precious human lives that are lost to road crashes. In the proposed system, we are designing and deploying a system that not only avoids accidents but also to take action accordingly.

B. Aim of the Work

This research aims at dealing with the issues that cause fatal crashes and also integrates measures to ensure safety. Life without transportation is impossible to imagine; it makes far off places easy to reach and greatly reduces the travel time. But the problems which surface due to the ever-increasing number of vehicles on the road cannot be ignored. The project aims to eradicate a few of the major reasons of car crashes and also aims to integrate post-crash measures. The reasons for automotive accidents focused here in this project are

- Nonchalant attitude towards the use of seat belts.
- Driving under the influence of alcohol.
- Distracted driving due to drowsiness.

The post-accident measure incorporated in the project is

- Intimation to the near and dear ones of the occurrence.

C. Objectives

The proposed project aims to achieve the following:

- Switch on the ignition only if the seat belts are locked in.
- Deploy a gas sensor to make sure that driver is not drunk. If the driver is not drunk, only then will the engine ignite.

- To ensure the driver is not drowsy, eye-blink sensors are deployed in the automobile.
- To circumvent a crash, a proximity sensor is deployed to discover the interruption in front of the automobile on the path.
- To ensure post-crash safety an alert system is deployed which makes use of a GPS system to attain the geographical location of the crashed vehicle and it is sent to a responsible and authorized individual. The accident is detected with the use of a vibration sensor.

D. Paper Organization

In the first section, an introduction has been provided to the whole project. All the fundamentals have been presented in which key modules of the project have been explained like the aims and objectives of the implemented system along with the motivation for choosing this project title. The second section of this project report reviews the literature surveys that have been performed to provide the basis for the implementation that is being performed. Equivalent and competing approaches that exist and have been worked upon are examined, recorded, and contrasted with the techniques and methods being implemented in this paper. These methods are further verified for any dichotomy that may be prevalent in their system. Further, methods are integrated to overcome those gaps. Starting from the third section, the technical aspect of the project is addressed. The basic framework and architecture of the methods are incorporated that will be realized in the building of smart vehicle safety and security systems. This is explained with the help of text and diagrams and flow charts. This helps in the step by-step visualization and organization of the project. The methodology of the project implementation is studied in greater depths in Section 4. Point-by-point software and hardware constraints and requirements to be met to accomplish the obvious building guidelines are further enrolled and comprehended in detail. The final section comprises of the conclusions. The same has been used to supply the basis for the brief of what has been done and further work scope of the project is discussed to provide a summary.

II. LITERATURE SURVEY

A. Survey of the Existing Models/Work

For Pannu et al. [1], the emphasis is on making a monocular vision, self-sufficient auto model utilizing Raspberry Pi as a handling chip [6]. A high-definition camera alongside an ultrasonic sensor was utilized to give fundamental information from this present reality to the automobile. The automobile is fit for achieving the given goal securely and insight fully in this manner avoiding the danger of human mistakes. Numerous current calculations like path identification and impediment location are consolidated to give vital control to the auto. The paper undertakes the implementation of the system using Raspberry Pi, by the ethicalness of its processor. Kumar et al. [2] proposed the design and development of an accelerometer based system for driver safety. This framework is structured by using Raspberry Pi (ARM11) for quickly accessing the control and accelerometer for event discovery. If any event occurs the message is sent to the authorized personnel so they can take quick and immediate response to save the lives and abate the

harms. The system only incorporates one module ignoring the other fatal causes thus making the proposed model incompetent and incomplete.

Sumit et al. [3] proposed a compelling strategy for the crash evasion arrangement of a vehicle to identify the hindrances present in the front and blind spot of the vehicle. The driver is alarmed with the help of a buzzer and an LED sign, as the distance between vehicle and obstacle reduces and is reflected on a display board. The ultrasonic sensor identifies the state of the object if it is moving or is stationary with respect to the vehicle. This system is valuable for discovering vehicles, bicycles, motorcycles, and pedestrians that cross by the lateral side of the automobile. The paper executes the proposed system using Raspberry Pi as the microcomputer but it limits out-of-the-box performance.

Mohamad et al. [4] proposed a proficient vehicle collision aversion framework inserted with an alcohol detector. This system has the capability of making the driver alert regarding the amount of alcohol consumed and depicting the same on an LCD screen. In addition it generates a warning using a buzzer to make the driver mindful of his or her own particular situation and to fag others in the encompassing zone [5]. The security segment proposed by this framework is the driver in an unusually abnormal state of tipsiness isn't allowed to drive an automobile as the start framework will be shut down. This method works in a way to intimidate the driver about his own condition, which is ironic because the person won't be mindful to take any action against it. The idea is novel but practically it is not workable.

B. Summary/Gaps Identified in the Survey

The current system showcases a mechanism for receiving the geographical coordinates of the automobile during a crash. This existent framework additionally provides a means of discovery of pre-crash with an object. But it does not target on the intensions that cause these fatal accidents. It does not focus on the crashes that are caused by drunk driving with the help of an alcohol/gas sensor and neither the negligence of use of seat belts.

Also these framework don't guarantee if the driver is wide awake or feeling drowsy. There is no use of eye-blink sensor for the same reason. Additionally, the current framework requires manual involvement. However, the proposed framework works on the shortcomings of the current work and is completely mechanized.

III. OVERVIEW OF THE PROPOSED SYSTEM

A. Introduction and Related Concepts

The proposed system utilizes an embedded system based on the Internet of Things and the Global System for Mobile Communication (GSM)[7] To avoid an accident, when the system is initiated, the seat belt is checked using a pressure sensor. If the driver is not wearing a seat belt, the engine is turned off. Then the alcohol sensor comes into play and checks for alcohol consumption, and if positive the engine is turned off. After these two main tasks [8], three things – tiredness, collision, and obstacles – are checked using the eye-blink sensor, vibration sensor [9], and the infrared (IR) sensor

[10], respectively. If there is a collision and the vibration sensor is active, then there is a message sent to the contact mentioned. If there is any obstacle present, then the buzzer beeps to tell the driver [11]. If the driver is feeling sleepy or drowsy, the eye blink sensor detects it and switches off the engine [12].

- 1) The system utilizes GSM technology for the communication of code pattern to transmit location coordinates.
- 2) The system is Arduino Uno based.
- 3) The system should be able to communicate even from physically far off distances.
- 4) The system uses an IR sensor[13], vibration sensor, alcohol sensor, eye-blink sensor, and pressure sensor [14].
- 5) To practically put together all the components and execute them, the composition of various sensing devices in our system is as shown in Figure 1

B. Proposed System Model

The software development system model that best suits this project and aligns itself with the needs of the given project is the Agile development model. Figure 2 is a block diagram, depicting the steps concerned in implementing the Agile development model.

In the Agile development model the entire requirement set is broken into numerous builds (Figure 3 and Figure 4). Various development stages take place here, making the development cycle a “multi-step waterfall” cycle. Cycles are split into tinier portions, making the modules easier to manage and implement. Every module goes through the planning, requirements analysis, design, implementation or building, and testing stages. A running version of the system is delivered at the end of the first iteration, so we get a working model early on during the product development cycle. Each iteration releases a model with added modules integrating more functions to the last release.

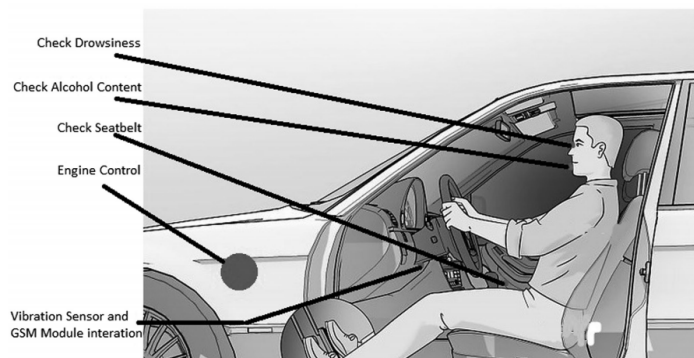


Fig. 1. System Design Implementation.

This process goes on until the complete system is developed. These iterations are repeated in a loop till an end version of the system is refined and is the expected outcome is obtained.

- As the project deploys a real-time checking and a monitoring system, the outputs produced are further

used to take the necessary actions and are thus fed back to the code to give an appropriate action for the further events.

- This process is redundant and cyclic in nature which is implemented whenever a driver enters the automobile.

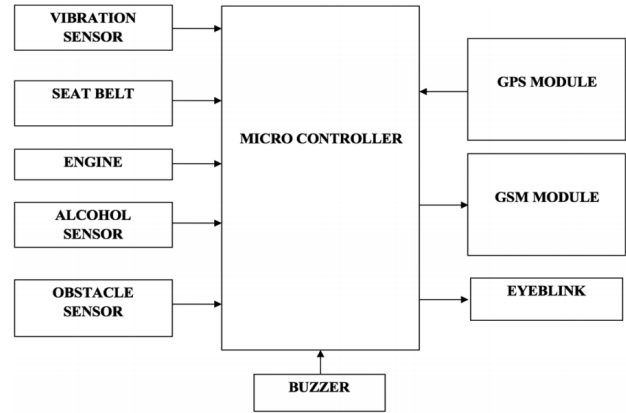


Fig. 2. Block Diagram.

IV. PROPOSED SYSTEM ANALYSIS AND DESIGN

A. Requirement Analysis

The arrangement of the idea to implement the functional requirements is elucidated under the heading system design. The arrangement of the idea to implement the non-functional requirements is elucidated in the system architecture section of this project report. The imperative functional requirements of this project's objective accomplishment are:

- 1) The automotive system should have the capability to determine whether the seat belt is put on or not by the driver.
- 2) The implemented system should have the capability to determine whether alcohol has been consumed by the driver or not.
- 3) The automotive system should have the capability to determine the mental awareness of the driver in terms of if he is feeling sleepy.
- 4) The automotive system should have the capability to check whether the vehicle is not coming too close to the vehicle in front.
- 5) The automotive system should have the capability to determine whether an accident has already taken place and thus should have the capability of sending the location coordinates of accident to a responsible person with the help of GSM technology.

This project deals with problems which cause accidents and attempts to ensure safety. This project addresses various reasons that lead to fatal accidents. Roads are unpredictable and at every turn of the road can be fatal accidents present and one cannot rely on the driving sense of other drivers and the pedestrians. One needs to be self-aware of the environment and the vehicles around. The driver should take all the precautions and be mindful of the people on the road as well because every life has value. Common reasons for accidents are the

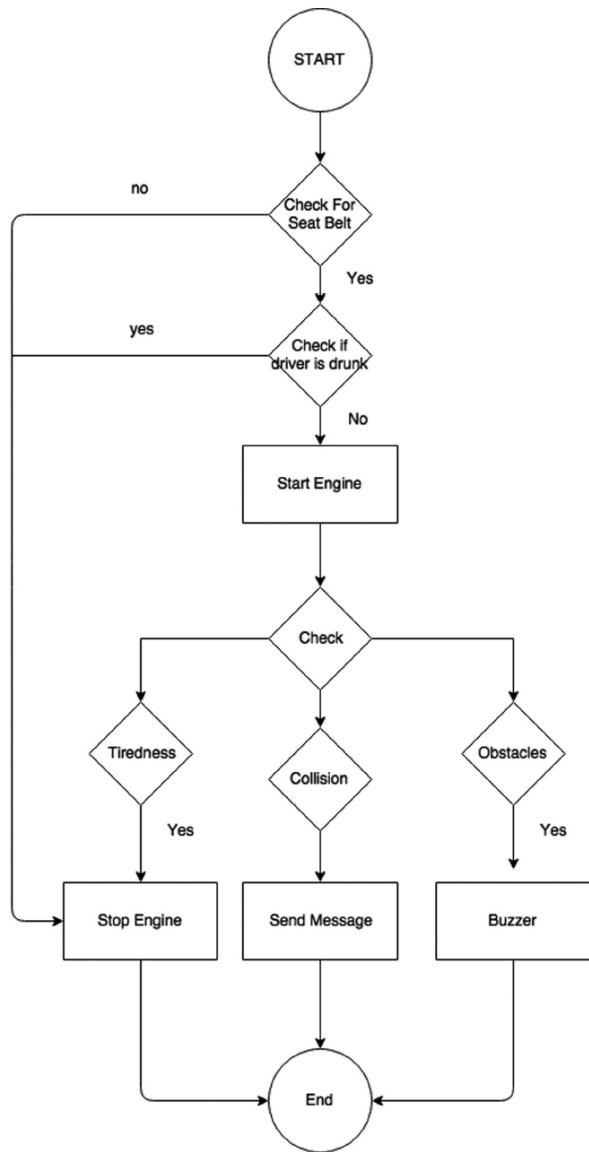


Fig. 3. UML Diagram (Activity Diagram).

lack of concentration of the driver on the road because of some distraction or because of lack of sleep of the driver. The product aims to provide the following functions:

- External forces should not result in the damaging of the system.
- The framework must be able to explicitly identify and discover problems related to the components.
- The issue detected should be reported back to the system.

The assumptions and dependencies are established in the beginning itself to give us a lucid understanding of the implementation of the product:

- Need of an appropriate GPS module to deliver exact geographical location coordinates.

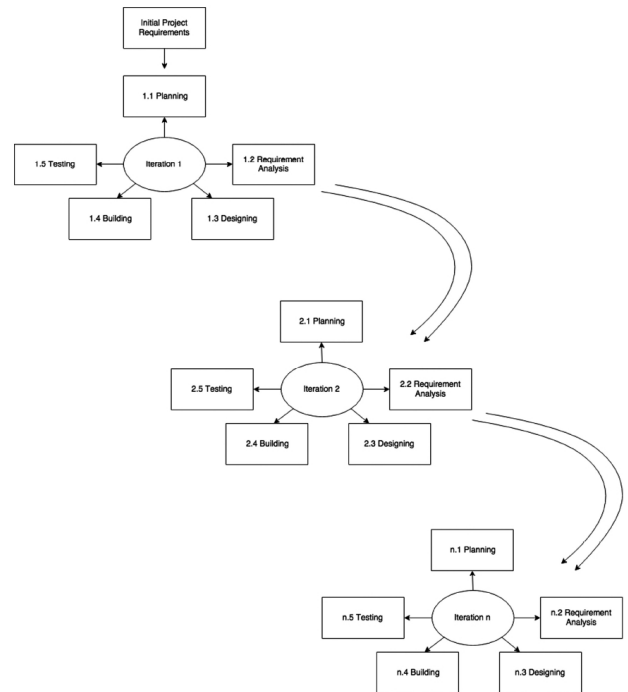


Fig. 4. Agile Model.

- The driver should be wearing the spectacles eye gear integrated with the eye-blink sensor.
- The system should always be connected to Internet.
- Proper placement of numerous proximity sensors can be added.

B. System Requirements

The output of the Smart Vehicle Security and Safety System is heavily dependent on Android application. The hardware components utilized for the project titled Smart Vehicle Security and Safety System are as follows. Arduino Uno board: The project utilizes Arduino Uno as the microcontroller. All the sensor components are attached and soldered to this microcontroller board and the microcontroller then takes the input and computes to give an appropriate output (Figure 5). Global vibration sensor: This project utilizes the vibration sensor to sense the accident and the crash of the automobile. This input received by the sensor is given

to the microcontroller Arduino Uno board that further utilizes the input to give a specific output (Figure 6). Alcohol/gas sensor: This project utilizes the alcohol sensor to sense the alcohol content in breath. This input received by the sensor is given to the microcontroller Arduino Uno board that further utilizes the input to give a specific output (Figure 7). Eye-blink sensor: This project utilizes the eye-blink sensor to sense the tiredness of the driver. This input received by the sensor is given to the microcontroller Arduino Uno board that further utilizes the input to give a specific output (Figure 8). Buzzer: This project utilizes the buzzer that signals and alerts the driver and the surroundings. The output is sent to the buzzer by the Arduino Uno board according to the computation (Figure 9). GPS module: This project utilizes

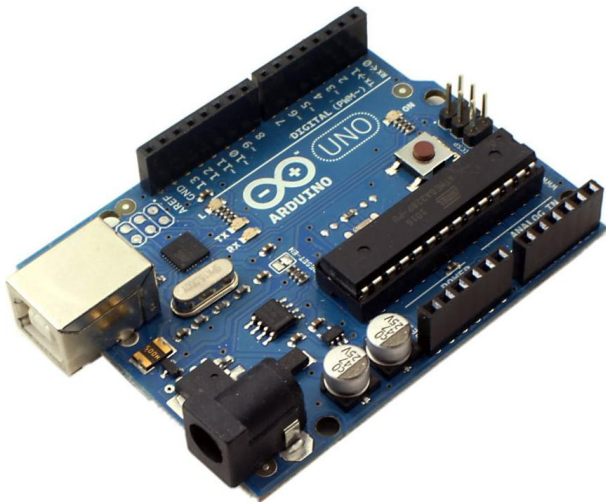


Fig. 5. Arduino Uno.

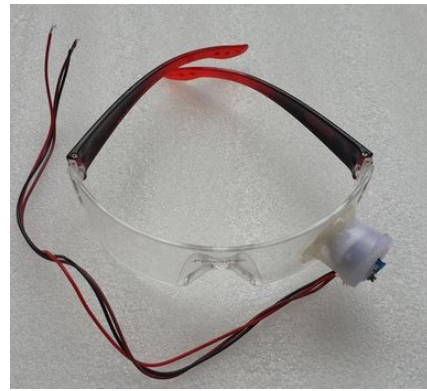


Fig. 8. Eye-blink Sensor.



Fig. 9. Buzzer.



Fig. 6. Vibration sensor.

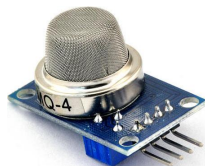


Fig. 7. Gas Sensor.



Fig. 10. GSM Module.

the GPS module to track the coordinates of the location of the where the project is present. This input received by the sensor is given to the microcontroller Arduino Uno board that further utilizes the input to give a specific output (Figure 10). GSM module: This project utilizes the GSM module to communicate the coordinates of the location as detected by the GPS module. This input received by the sensor is given to

the microcontroller Arduino Uno board that further utilizes the input to give a specific output (Figure 10). This instructional exercise will disclose how to interface a GSM modem with Toradex modules (Figure 11).

V. RESULTS AND DISCUSSION

A. Experimental Results

The experimental results show that the proposed model gives us a better result as compared with other available devices. The output of the force sensitivity sensor is shown in Figure 12. This figure shows that the output provides more values as per the increase in time. Figure 13 shows the serial monitor of the vibration sensor.

Whereas, Figure 14 and Figure 15 display the percentage of crashes due to fatigue and causes of crashes, respectively.

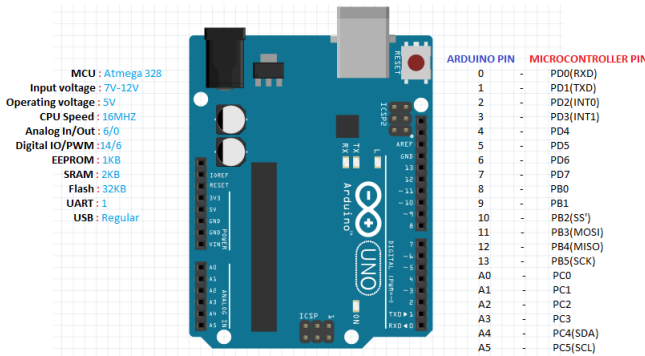


Fig. 11. Arduino Mapping.

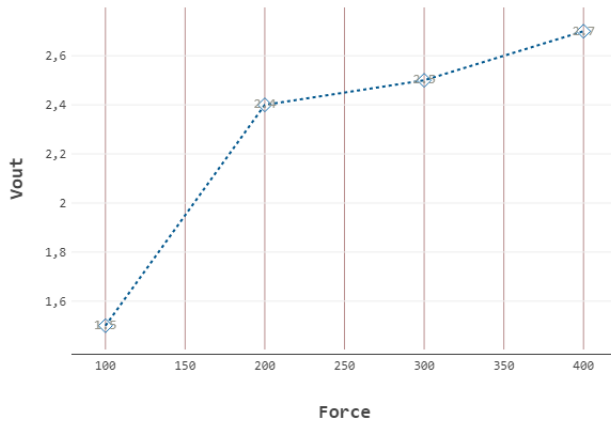


Fig. 12. Force Sensitivity Sensor.

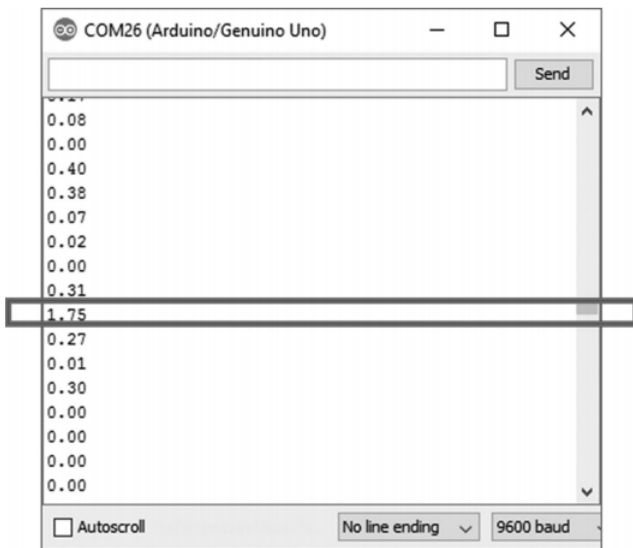


Fig. 13. Serial Monitor for Vibration Sensor.

B. Final Output of the Research and Conclusion

A competent Smart Vehicle Security and Safety System integrated with a pressure sensor, eye-blink sensor, alcohol

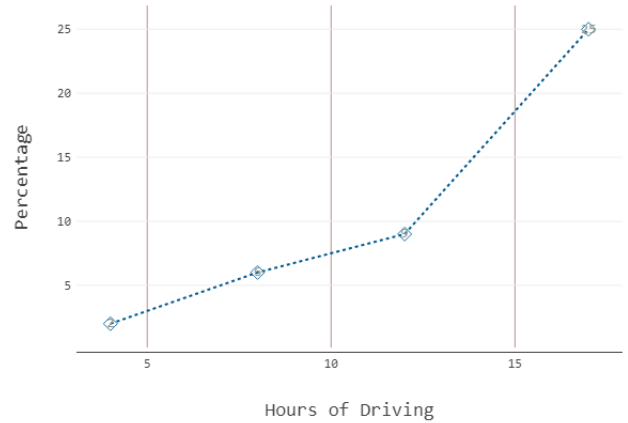


Fig. 14. Percentage of Crashes Due to Fatigue.

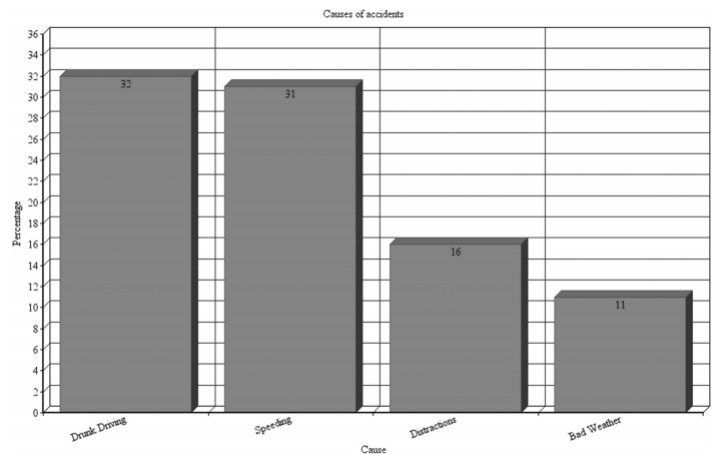


Fig. 15. Causes of Crashes.

sensor, proximity sensor, and vibration sensor (Figure 16) using the concept of GPS and GSM has been implemented. The sensors are integrated with the Arduino board. Areas with outreach problems that experience bad network connectivity or in remote areas with no network connectivity available can be an issue.

This can in turn lead to the accident intimidation text not being sent to the specified number. The proposed and thus implemented system can be enhanced and modified by adding concepts of technology such as big data and GPS to study the thus collected data to understand and read the patterns associated with the crashes. The same system can be modified accordingly and implemented for two Wheeler's. Further, the location of the crash can be sent to an ambulance as well for quick medical response and attention.

REFERENCES

[1] G. S. Pannu, M. D. Ansari, and P. Gupta, "design and implementation of autonomous car using raspberry pi." international journal of computer applications 113," no., vol. 9, 2015.

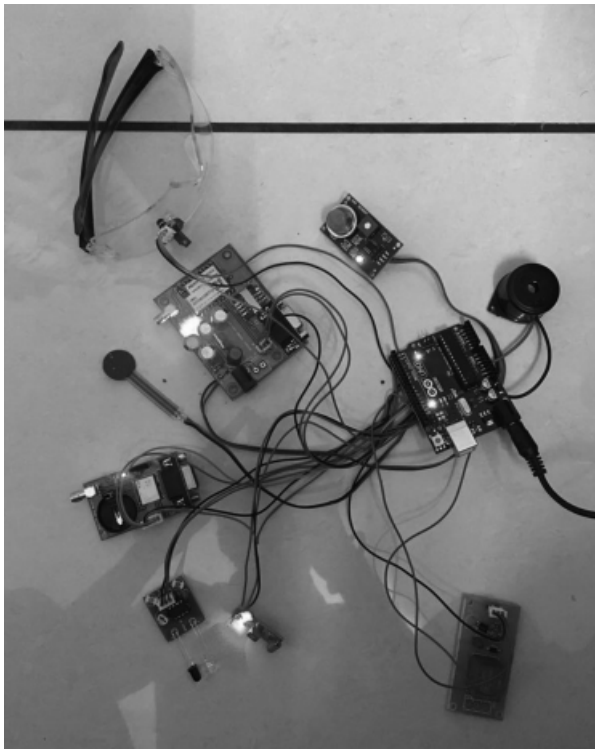


Fig. 16. Hardware Prototype.

- [2] V. N. Kumar, V. S. Reddy, and L. P. Sree, "design and development of accelerometer based system for driver safety," *international journal of science*, *Engineering and Technology Research (IJSETR)*, vol. 3, p. 12, 2014.
- [3] C. Hahn, S. Feld, and H. Schroter, "Predictive collision management for time and risk dependent path planning," in *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 405–408. [Online]. Available: <https://doi.org/10.1145/3397536.3422252>
- [4] G. N. A. H. Yar, A.-B. Noor-ul Hassan, and H. Siddiqui, "Real-time shallow water image retrieval and enhancement for low-cost unmanned underwater vehicle using raspberry pi," in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, ser. SAC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1891–1899. [Online]. Available: <https://doi.org/10.1145/3412841.3442060>
- [5] A. F. B. A. de Oliveira and L. V. L. Filgueiras, "Developer assistance tools for creating native mobile applications accessible to visually impaired people: A systematic review," in *Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems*, ser. IHC 2018. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: <https://doi.org/10.1145/3274192.3274208>
- [6] S. L. Fong, D. C. W. Yung, F. Y. H. Ahmed, and A. Jamal, "Smart city bus application with quick response (qr) code payment," ser. ICSCA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 248–252. [Online]. Available: <https://doi.org/10.1145/3316615.3316718>
- [7] G. K. Gudur, A. Ramesh, and S. R., "A vision-based deep on-device intelligent bus stop recognition system," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, ser. UbiComp/ISWC '19 Adjunct. New York, NY, USA: Association for Computing Machinery, 2019, p. 963–968. [Online]. Available: <https://doi.org/10.1145/3341162.3349323>
- [8] D. Saha, M. Shinde, and S. Thadeshwar, "Iot based air quality monitoring system using wireless sensors deployed in public bus services," ser. ICC '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: <https://doi.org/10.1145/3018896.3025135>
- [9] M. Kumar, "R., and dr," *R. Senthil. Effective control of accidents using routing and tracking system with integrated network of sensors*, vol. 2, p. 4, 2013.
- [10] R. Liu, Z. Yin, W. Jiang, and T. He, "Wibeacon: Expanding ble location-based services via wifi," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 83–96. [Online]. Available: <https://doi.org/10.1145/3447993.3448615>
- [11] J. J. T. Dai, X. Bai, and Z. Shen, "Mobile phone based drunk driving detection pervasive computing technologies for healthcare. 2010, 4th international ieee conference," p, vol. 1, March 2010.
- [12] H. Chen, Y. Chiang, F. Chang, and H. Wang, "Toward real-time precise point positioning: Differential gps based on igs ultra rapid product. sice annual conference," *The Grand Hotel, Taipei, Taiwan, August*, vol. 18.
- [13] X. Liu, X. Xu, X. Chen, E. Mai, H. Y. Noh, P. Zhang, and L. Zhang, "Individualized calibration of industrial-grade gas sensors in air quality sensing system," ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: <https://doi.org/10.1145/3131672.3136998>
- [14] A. T. Duchowski, S. Jörg, T. N. Allen, I. Giannopoulos, and K. Krejtz, "Eye movement synthesis," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 147–154. [Online]. Available: <https://doi.org/10.1145/2857491.2857528>

Permissioned Blockchain: Securing Industrial IoT Environments

Samira Yeasmin¹, Adeel Baig²

Department of Computer Engineering
Al Yamamah University, Riyadh, Saudi Arabia

Abstract—With the significantly increased use of the Industrial Internet of Things (IIoT), it is believed that this technology will revolutionize industrial applications and infrastructures by connecting several industrial assets. But it is getting prone to many cyberattacks and security issues. The emerging security challenges of IIoT can have a devastating effect since it deals with mission and safety-critical systems. Thus, it becomes extremely important to address the security vulnerabilities and susceptibilities of this technology. Blockchain, being one of the most significant solutions to several technologies' security problems, can play a vital role in improving the security of IIoT. Therefore, this paper proposes to use a Hyperledger Fabric Blockchain-enabled IIoT that guarantees the security of the communication medium, data storage, access, and sharing between the IIoT devices and ensures to provide limited access to the authorized identities only. This system also monitors the user access and makes sure that the transactions are performed according to their roles defined by the Certificate Authority (CA) and Membership Service Provider (MSP). Moreover, this paper presents the findings on the implementation of the blockchain network and addresses the key challenges. It evaluates the performance of the proposed network and discusses the key areas to be improved. Finally, the paper describes the benefits of the permissioned blockchain for IIoT and presents a future direction for further research and study.

Keywords—Industrial Internet of Things; IIoT; permissioned blockchain; hyperledger fabric; information security; device communication; data sharing; access control

I. INTRODUCTION

Connecting numerous industrial devices to share information and make important business decisions, the Industrial Internet of Things (IIoT) is the core to aim and realize intelligent industrial manufacturing and production. It is mainly used in the mission and safety-critical systems [1] that allow making better decisions to improve the systems' efficiency. Although IIoT is believed to be competent to enhance industrial assets and digitally transform the industrial infrastructures, the centralized network creates a vulnerable environment [1]. The heterogeneous network of IIoT devices increases cyber threats including insecure IoT gateways and MQTT protocol, insecure cyber-physical systems (CPS), and SCADA [1]. Therefore, it becomes extremely important to address the security challenges and take appropriate measures to improve them.

Blockchain technology is gaining popularity in both industrial and academic research fields, showing promising results in solving the security challenges of the arising

technologies. The use of blockchain in the IIoT field will improve the cyber threats and safeguard it from malicious activities that might occur during the communication between the IIoT devices. Blockchain is a Distributed Ledger Technology (DLT) that was primarily used for storing transaction information. It is a distributed network of multiple computers connected in a peer-to-peer network. The use of cryptographic security makes it suitable to secure the device communication of the IIoT network.

Taking this into account, this paper proposes to use a permissioned blockchain to secure the device communication, address and improve the security vulnerabilities of IIoT. The unique features of the permissioned blockchain including identity management and restricted user access enables only authorized parties will participate in performing transactions and device communication.

Though various studies show the implementation of Blockchain in improving IIoT security, the use of a Permissioned Blockchain is not heavily researched and implemented. The application of a Public Blockchain already exists in many forms including lightweight authentication mechanism, federated learning approach, and use of different types of encryption [2]–[9]. However, it leads to lower throughput, higher latency and resource utilization. Despite the vast literature, the implementation of Hyperledger Fabric Blockchain in securing the IIoT device communication has not been well recognized. The true benefit of a Permissioned Blockchain, the use of a Certificate Authority (CA) to issue certificates, and Membership Service Provider (MSP) to define an access control mechanism is yet to be unleashed. Validation and verification of each transaction, communication accessibility and transaction invocation to only allowed participants ensures higher throughput and lower latency, leading to an improved and secured IIoT environment.

To summarize, the main contribution of our paper is as follows: (1) this paper provides a background study on IIoT and blockchain and a comparison analysis between different open-source blockchain platforms; (2) a detailed discussion is made to state the security issues of IIoT device communication and implements the proposed idea to use a Permissioned Blockchain called Hyperledger Fabric; (3) the performance of the blockchain is analyzed and evaluated, and finally, (4) directions for future works are identified.

This paper's organization is structured as follows: Section II focuses on the background and Section III describes the need to improve the security of IIoT device communication. In

Section IV, some of the existing literature has been discussed. The proposal and a detailed discussion have been made in Sections V and VI. Section VII presents the hypothesis. The system's implementation is discussed in Section VIII. Section IX is dedicated to the evaluation of the implemented solution. Finally, Sections X and XI present the future study and conclusion.

II. BACKGROUND

A. Industrial Internet of Things (IIoT)

As a subset of IoT, IIoT is defined as a connection between machines, computers, and people that enables intelligent industrial operation [10] by collecting data through wireless sensor networks, communication protocols, and internet infrastructure. This data is analyzed to produce important results that help in faster and more accurate business decisions. Although IIoT provides many benefits, it faces some challenges, especially in security, and privacy [11].

While IoT and IIoT might sound similar, Table I shows the comparison between them where the main difference is mainly in the area of interest, network, connectivity, and performance.

B. Blockchain

Blockchain is the foundation of cryptocurrency transactions and is a distributed database providing transparent, secure, and fast transactions. It is a chain of data blocks containing a time-stamp for each block [13]. It enables different parties to form and maintain consensus without an intermediary. Blockchain is decentralized, immutable, anonymous, and cryptographically sealed [14]. Table II presents the comparison between the three types of blockchain, having the main difference in network type, consensus, and read-write (RW) permissions.

C. Blockchain Platforms

Many blockchain platforms allow building decentralized applications including the open-source blockchain platforms presented in Table III. The main difference is in the type of blockchain network that also defines the transaction visibility such as public, permissioned, or private. Only Hyperledger uses a pluggable consensus protocol. Moreover, a higher throughput generates a lower latency, increasing energy and computational costs.

D. Hyperledger Fabric

Hyperledger Fabric is a permissioned blockchain where all the participants have a registered id, and all transactions are private and confidential [28], also are authenticated, authorized. It implements a distributed ledger platform to run Chaincode [28], delivering a high degree of resiliency, flexibility, confidentiality. It supports a pluggable consensus protocol [29].

1) Key Components of Hyperledger Fabric

a) *Certificate Authority (CA)*: An CA is responsible for creating, managing, and issuing certificates to different network actors by providing them with a pair of public and private keys, restricting user access [30]. The certificates are

digitally signed and bind together with the actor's public key. It issues a root and enrollment certificate and allocates a transaction certificate to each authorized member [31].

b) *Membership Service Provider (MSP)*: It provides membership permission based on the certificates and delivers services such as identity validation, user registration, and authentication, also assign appropriate permission. It decides if the user will be a peer, admin, client, orderer, or member [30]. This component is installed on each channel peer to ensure transactions are authenticated [28]. After the CA provides a key pair and the transactions are signed using a public key, MSP verifies the transaction [30].

c) *Peers [30]*: The peer nodes host ledger and smart contracts, encapsulating shared processes and information.

- **Endorser/Endorsing Peer**: This peer validates the transaction and executes the Chaincode without updating the ledger. In the end, the endorser might approve or reject the transaction.
- **Orderer Peer**: It does transaction ordering, creating, and delivering new block to all the peers, eliminating bottlenecks.
- **Anchor Peer**: When a configuration block has updates, this peer broadcasts the updates to the rest of the peers. Anchor peers are discoverable and can be communicated by all the other peers of the network.

TABLE I. COMPARISON BETWEEN IOT AND IIOT

Area	IoT	IIoT
Focus	Consumer-level devices	Mission or safety-critical systems
Service [11]	Human-centered	Machine-centered
Architecture	3 or 5 layers [12]	3 layers [10]
Communication	Business-to-Consumer	Business-to-Business
Used in [11]	New devices & standards	Existing devices & standards
Connectivity [11]	Ad hoc	Structured
Volume of data [11]	Medium to high	High to very high
Scalability	Used in low scale network	Used in large scale network
Latency & Speed	Utility centric	High speed & minimum latency is required

TABLE II. COMPARISON OF THREE TYPES OF BLOCKCHAIN

	Public	Consortium	Private
Network	Decentralized	Partially centralized	Centralized
Consensus	Permissionless	Permissioned	Permissioned
RW	Public	Public/Permissioned	Permissioned
Example	Bitcoin, Ethereum, Litecoin	Quorum, Hyperledger, Corda	Bankchain

TABLE III. COMPARISON BETWEEN DIFFERENT OPEN-SOURCE BLOCKCHAIN PLATFORMS

Ethereum	
Blockchain Type & Network	Public/Private & Decentralized [14]
Consensus Algorithm	PoW
Cryptocurrency	Ether
Smart Contract	Written in Solidity
Vulnerability to attacks	51% attack [15]
Data Confidentiality	No
User Authentication	Digital Signature
Throughput & Latency	6-7 TPS [16] & 15-20 sec
Energy & Computational Cost	High [16]
Hyperledger	
Blockchain Type & Network	Consortium/Partially centralized [14]
Consensus Algorithm	No consensus or Pluggable consensus or Practical Byzantine Fault Tolerance
Cryptocurrency	No native cryptocurrency
Smart Contract	Written in Go, Java, Node.js
Vulnerability to attacks [17]	>1/3 faulty nodes [15] & DoS attack
Data Confidentiality	Yes
User Authentication	Based on enrolment certificates
Throughput & Latency	>1,000 TPS [18] & Less than Ethereum
Energy & Computational Cost	Low [15]
Corda	
Blockchain Type & Network	Consortium & Decentralized
Consensus Algorithm [19], [20]	Pluggable consensus, Validity Consensus & Uniqueness Consensus
Cryptocurrency	No native cryptocurrency
Smart Contract	Written in Kotlin and Java [20]
Vulnerability to attacks	Denial-of-state (DoSt) attack [21]
Data Confidentiality	Yes
User Authentication	Digital signatures
Throughput & Latency	600 TPS [22] & Low
Energy & Computational Cost	High [23]
Openchain	
Blockchain Type & Network	Private & Decentralized
Consensus Algorithm	Partitioned Consensus [24] and PoA
Cryptocurrency	No native cryptocurrency
Smart Contract	No [25]
Vulnerability to attacks	-----
Data Confidentiality	Yes
User Authentication	Digital signatures
Throughput & Latency	1000 TPS & Low
Energy & Computational Cost	High
IOTA	
Blockchain Type & Network	Public [15] & Partially centralized
Consensus Algorithm	Tip Selection Algorithm
Cryptocurrency	mIOTA
Smart Contract	No [21]
Vulnerability to attacks	34% attack [15]
Data Confidentiality	No
User Authentication	Digital signatures
Throughput [15] & Latency	7-12 TPS & Varies from mins to hours

Energy & Computational Cost	Low [15]
Ripple	
Blockchain Type & Network	Consortium & Decentralized/Centralized
Consensus Algorithm [26]	Ripple Consensus Algorithm (RPCA)
Cryptocurrency	Ripple (XRP)
Smart Contract	No [25]
Vulnerability to attacks	DoS & Theft attack [25]
Data Confidentiality	Yes
User Authentication	Digital signatures
Throughput & Latency	1,500 TPS [27] & Low
Energy & Computational Cost	Low [27]

III. PROBLEM STATEMENT

While IIoT devices can improve efficiency, it also comes with potential cybersecurity challenges. Since all the devices are connected, security becomes the prime concern while implementing it. The centralized nature of IIoT devices makes it open to different cyberattacks since compromising one single point can infect and destabilize the whole network. IIoT communication is transparent between the stakeholders and it makes the security problem worse as it becomes susceptible to different kinds of cyber-threats such as Man-in-the-Middle (MITM) and Denial-of-Service (DoS) attacks [32]. Communication between IIoT devices needs to be secured since they generate, process, and exchange a huge amount of data that is related to the mission and safety-critical infrastructures. A data breach may happen while sharing or transmitting data. Since IIoT devices are being used in mission and safety-critical systems, a key issue is to protect these valuable and sensitive data. Therefore, there is a need to address and improve data security by securing the communication between IIoT devices. Using Hyperledger Fabric Blockchain, the following questions were investigated: (a) Can Hyperledger Fabric provide better confidentiality and integrity compared to current approaches? (b) Does Hyperledger Fabric CA and MSP ensure access control and provide trust between devices? (c) Can a permissioned blockchain be integrated with IIoT to achieve higher throughput, lower latency, and better resource utilization?

IV. LITERATURE REVIEW

Authors in [2] proposed a lightweight authentication mechanism for industrial device communication using simple hashing functions to complete device authentication. In [3], authors studied the CLS scheme of [4] and presented its vulnerability towards public key replacement attacks to achieve data authenticity in IIoT. A PoW credit-based consensus algorithm is used for IIoT devices in [5]. The DAG-blockchain-based architecture included device authorization and proposed wireless sensors to act as light nodes having Private Key (PK) and Secret Key (SK) to sign transactions, and gateway and manager to act as full nodes having PK hardcoded in gateways.

The author in [6] proposed using a dynamic secret sharing mechanism in the IIoT data transmission technique using power blockchain. It included users to submit transactions and issue certificates (TCerts). Paper [7] proposed a multi-party

data-sharing model using a permissioned blockchain where only registered participants could share or access data and run a consensus algorithm named Proof of training quality (PoQ) using a federated learning approach. In [8], the authors combined supply chain, IIoT, and blockchain to securely share data using attribute-based encryption. The mechanism included the registration of nodes and the definition of user roles according to the smart contract or signature provided by the admin. Authors in [9] proposed a blockchain-based IIoT architecture to improve processing power, security and privacy. Whitelist and blacklist mechanisms were used to restrict access and allow transactions via PoW.

V. PROPOSED IDEA

A huge amount of data is generated and shared between the IIoT devices, including sensitive information. Therefore, a secure communication medium is required to enhance data security and privacy. The proposed idea is to use a permissioned blockchain that will only allow authorized members to access the network. It will permit only one or more nodes to work together to control and restrict access of members of the chain network. As a permissioned blockchain, Hyperledger Fabric can be used that allows communication only between the authorized members. The network's goal is to enforce a trusted device communication between multiple parties connected through IIoT. All the other members outside the network will be considered malicious to secure it from cyberattacks. Each member of the network is responsible for setting up their peers authorized by a CA. An MSP will allow permissions based on the CA and define access control rules. Endorsing Peer will execute transactions and Anchor Peer will update other peers. Also, Orderer Peer will create and deliver new blocks. It is important to mention here that the transactions in the entire blockchain network include data storage, access, sharing, and monitoring by the organizations' admins.

While communication happens between two different IIoT devices, only allowed participants will perform the transactions based on their roles defined by MSP. The communication medium will be secured enough as the peers will verify and validate each transaction before adding it to the network's ledger.

VI. HYPERLEDGER FABRIC BLOCKCHAIN-ENABLED IIOT

The use of a Hyperledger Fabric blockchain ensures a secured environment where only permissioned organizations can perform transactions. One of its main features includes

organization members acting as participants. But before deciding which members will participate in the blockchain, a CA creates identities and MSP defines user access roles.

A. CA

The built-in Fabric CA [30] plays a vital role in securing the device communication of IIoT devices. A server and a client component make up the Fabric CA. It is used to create a new root CA that works as per the requirement of the system. The root CA creates an intermediate CA that generates certificates to the identities. The same database is shared among all the CA servers to keep track of identities and certificates. To register a new identity, the registrar will need an attribute along with a value because the new identity's affiliation and the registrar's attribute must be equal. If these conditions are met, the CA creates an identity and provides a keypair that consists of a public and private key. The public key is used to sign a transaction since the private key cannot be shared publicly. The MSP component ensures the verification of the transactions.

B. MSP

After registering a new identity on the CA, the MSP defines the role based on the certificate. The roles include a peer, admin, client, orderer, or member. Since the attribute and value are used to register a new identity on CA, it becomes easy for the MSP to decide a particular role for the participant. This allows an access control mechanism in the system. According to MSP, only 'client' identities can invoke transactions [30]. Whereas admins handle administrative tasks and peers take care of the transactions and ordering. While invoking a new transaction, the clients use their public key to sign the transaction and if it matches with the private key, it is added to the transaction. The MSP ordering service contains all the public keys of the clients. The verification process includes the MSP to check if the public key matches the public key it has. If verified, it is sent to the endorser peer for further processing.

C. Endorser Peer

After receiving the transaction invocation request, it checks the certificate detail and role of the transaction requester. The Chaincode is executed in this phase, following the endorser peer to decide the execution of the transaction. It is important to note that, only the endorser peer is having the Chaincode, therefore, it does not need to be installed on every node of the blockchain. Thus, it increases the scalability of the blockchain network. After deciding the validity of the transaction, the endorsement response, including the RW set, is sent to the client who invoked the transaction. If approved, the client sends the approved transaction to the orderer peer to process it.

D. Orderer Peer

The orderer peer, the central communication channel, adds the transaction into a block. Kafka is an ordering mechanism that helps in having a fault-tolerance solution for transaction ordering. To provide consistency across the whole network, the orderer peer orders all the transactions sequentially and prevents the double-spending attack. After adding the transaction into a block, it is forwarded to the organization's members to commit to the ledger. However, the verification

policy is run here again to verify if the transaction has been endorsed according to the Chaincode endorsement policy.

E. Anchor Peer

The anchor peer updates and notifies the other peers about the inclusion of the new block to the ledger. The local ledger is updated with the newly added block. It helps in maintaining synchronization across the whole network.

Fig. 1 explains the process of adding a new identity to the Hyperledger Fabric Blockchain network and invoking a transaction.

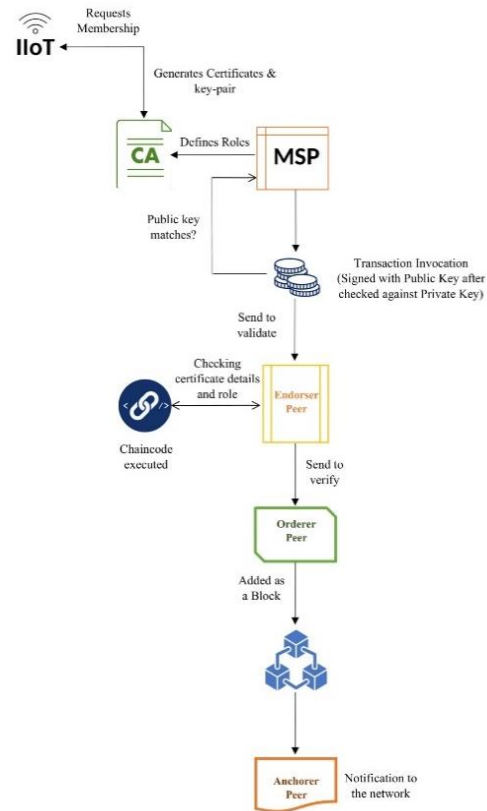


Fig. 1. Process of Hyperledger Fabric Blockchain-enabled IIoT.

VII. HYPOTHESIS

The Hyperledger Fabric Blockchain improves and enhances the IIoT network's security through the restricted participation of members of the organizations. Whenever a new organization wants to join the network, the CA generates a certificate to verify access. Moreover, the newly added organization's configuration settings and access control mechanism are defined by the MSP enabling the network to be secured from unauthorized access and transaction performance. Since the blockchain and transaction execution is restricted by the defined access control mechanism, the proposal ensures data integrity and confidentiality across the whole network. Every transaction is validated by the endorsement policy of the network, allowing only authorized parties to get involved in this process. This permissioned network is well suited for the IIoT environment to ensure data is stored, accessed, and shared only between trusted parties to achieve privacy and security.

VIII. IMPLEMENTATION

The implementation details of the proposed Hyperledger Fabric Blockchain-enabled IIoT include the addition of IIoT devices to the existing channel of the open-source project and performing transactions in a secured manner as hypothesized. In order to implement the proposed system, the open-source project Hyperledger Fabric blockchain has been chosen since it meets the requirements of our system such as CA, MSP, the three types of peer components. As Hyperledger Fabric is a permissioned blockchain and adds a level of data security and privacy, it suits the implementation requirements of this system.

As hypothesized, a CA is generated for all the Hyperledger Fabric Blockchain-enabled IIoT devices, allowing MSP to define user roles. This step allows an access control mechanism for the devices and secures the network from unauthorized access and transaction performance. Moreover, endorsers verify the transactions according to the user definition and authenticate the signatures with the CA. Therefore, CA generated certificates, the access control mechanism, and configuration updates defined by the MSP can provide a privacy protection layer across the whole network and prevent malicious network intrusion.

A. Setting up the Environment

The blockchain network has been implemented on Hyperledger Fabric v1.1.0. The experiments have been carried out on a Virtual Machine (VM). The VM acts as the IIoT environment where the Hyperledger Fabric Blockchain is configured, and IIoT devices are created. We selected VM as it gives the flexibility to implement the required configurations and contributes to our concept of a secure and limited resource IIoT environment. The hardware and software environments are described in Table IV.

TABLE IV. HARDWARE AND SOFTWARE ENVIRONMENT SPECIFICATIONS

Type	Environment	Specification
Hardware	CPU	Intel® Core™ i7-7500U CPU @ 2.70GHz
	Memory	6GB
	Hard Disk	20GB
Software	OS	Ubuntu 18.04 LTS
	docker	19.03.13
	docker-compose	1.17.1
	nodejs	8.10.0
	npm	5.3.0
	golang	1.10.4

The fabric network has been created with two organizations that are already provided by the open-source project. Both organizations consist of two endorsing peers and a CA. Each channel of the blockchain consists of a number of IIoT devices that can interact with each other through performing transactions and without any intermediaries.

B. Adding a New IIoT Device

To simulate the proposed system, three organizations act as three IIoT devices. It has been assumed that each organization

plays the role of an IIoT device and each device has two running peers. Org1, Org2, and Org3 represent the devices. The Hyperledger Fabric comes with two organizations already developed in the network, Org1 and Org2. Therefore, our simulation focuses on adding a new IIoT device, Org3 that requires generating certificates and configurations.

To add a new organization, the sub-directory “first-network” of the root directory “fabric-samples” has been used. After launching and bringing up the existing IIoT network using docker-compose, crypto materials and certificates were generated for the new IIoT device. Configuration files were also prepared for crypto-config and transactions. Also, configuration materials were generated for the new device. The org3-crypto.yaml file generates keys and certificates for the new IIoT device and creates two peers. Therefore, the artifacts of the Org3, IIoT device 3, configuration file consisted of an admin user certificate, a CA, an MSP, TLS certificates, two peers, and users. The configuration files are shown in Fig. 2.

To update the processes, a configuration tool named configtxlator has been used. It performs transactions and configures tasks, providing a stateless REST API without an SDK [30]. Using a Command Line Interface (CLI) helps in encoding and decoding between protobufs and JSON. To add the new IIoT device to the existing channel of the blockchain, this tool fetches a new configuration block and updates the information inside it. This step prevents repetition in the configuration changes and adds Org3MSP to the network. This procedure updated the existing config.json file to a modified_config.json file containing the new IIoT device's configurations.

To support the concept of security and write the configuration updates on the ledger, the Org1 and Org2 admin signs on peer0.org1 and peer0.org2. Then the Orderer processes the signatures and adds a new block to the network. Therefore, the block height is changed from 5 to 6, as shown in Fig. 3. In this way, the new IIoT device gets defined in the channel and is ready to become a part of it.

To join the channel, the peers of the device need to be up and running using docker-compose. The genesis block, the first block of the network, is copied in the CLI for the IIoT device specifying the environment variables. The ordering service can verify the new device by receiving a call and successfully adds it to the channel. The new device's signature is added to the call for service while sending it to Orderer for verification purposes. Otherwise, the ordering service rejects the call. Fig. 4 represents the proposal submitted by the new device to join the existing IIoT-blockchain channel. After acceptance from Orderer, the IIoT device is added to the channel.

```
sanirayeasmin@sanirayeasmin-VirtualBox:~/fabric-samples/first-network/org3-artifacts/crypto-config/peerOrganizations/org3.example.com$ ls  
ca msp peers tlsca users
```

Fig. 2. IIoT Device Configuration Files.

```
root@53cedfa3aa56:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer channel getinfo -c mychannel  
2020-10-09 09:28:00.403 UTC [channelCmd] InitCmdFactory -> INFO 001 Endorser and orderer connections initialized  
Blockchain info: {"height":6,"currentBlockHash":"gGAKWV5ZDw9cB3p531fTG2b/dv6UIzJPvWV40t5Z1V0=", "previousBlockHash":"qBd/uwIuvJ2LrA50k8hZfBNFGEHIX2CkMF1qNpZ8sI="}
```

Fig. 3. Block Height.

```
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chan
nel join -b mychannel.block
2020-10-10 09:12:08.171 UTC [channelCmd] InitCmdFactory -> INFO 001 Endorser an
d orderer connections initialized
2020-10-10 09:12:08.227 UTC [channelCmd] executeJoin -> INFO 002 Successfully s
ubmitted proposal to join channel
2020-10-10 09:12:08.227 UTC [main] main -> INFO 003 Exiting.....
```

Fig. 4. New IIoT Device Joining the Blockchain Channel.

C. Experimental Results

To perform some transactions, the Chaincode needs to be updated for all the peers of the devices so that Chaincode instantiation can be made. This step allows the newly added device's endorsement policy to be consistent with the rest of the devices, and it also ensures the newly added device is a valid member to endorse transaction invocation.

After specifying the endorsement policy and upgrading the Chaincode for the new IIoT device, some transactions are queried to evaluate the performance of the device communication. To begin with, a call is instantiated with a value of “a” to be 90 and “b” to be 210. The instantiation and endorsement policy is represented in Fig. 5. This allows the new device to perform transactions during the endorsement phase. Fig. 6 represents transaction execution, and it is performed between the peers of the devices. Two transactions are executed. The first one sends a value of 10 to move from “a” to “b”. Therefore, the value of “a” is 80, and “b” is 220 after performing this transaction.

Likewise, Fig. 7 shows one more transaction that is invoked with a value of 30 to be moved from “a” to “b”, making “a” to be 50 and “b” to be 250. This is how the devices and peers communicate with each other by performing transactions and following the endorsement policy.

```
root@bec29de5afd0:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode upgrade -o orderer.example.com:7050 --tls SCORE_PEER_TLS_ENABLED --cafile
$ORDERER_CA -C $CHANNEL_NAME -n mycc -v 2.0 -c '{"Args":["init","a","90","b",
"210"]}' -P "OR ('Org1MSP.peer','Org2MSP.peer','Org3MSP.peer')"
2020-10-10 09:16:55.757 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:16:55.757 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
2020-10-10 09:17:21.213 UTC [main] main -> INFO 003 Exiting.....
```

Fig. 5. Endorsement Policy and Chaincode Instantiation.

```
root@bec29de5afd0:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode upgrade -o orderer.example.com:7050 --tls SCORE_PEER_TLS_ENABLED --cafile
$ORDERER_CA -C $CHANNEL_NAME -n mycc -v 2.0 -c '{"Args":["init","a","90","b",
"210"]}' -P "OR ('Org1MSP.peer','Org2MSP.peer','Org3MSP.peer')"
2020-10-10 09:16:55.757 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:16:55.757 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
2020-10-10 09:17:21.213 UTC [main] main -> INFO 003 Exiting.....
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode query -C $CHANNEL_NAME -n mycc -c '{"Args":["query","a"]}'
2020-10-10 09:18:30.827 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:18:30.828 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
Query Result: 90
2020-10-10 09:18:54.976 UTC [main] main -> INFO 003 Exiting.....
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode invoke -o orderer.example.com:7050 --tls SCORE_PEER_TLS_ENABLED --cafile
$ORDERER_CA -C $CHANNEL_NAME -n mycc -c '{"Args":["invoke","a","b","10"]}'
2020-10-10 09:19:24.151 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:19:24.151 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
2020-10-10 09:19:24.169 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 003 C
haincode invoke successful. result: status:200
2020-10-10 09:19:24.170 UTC [main] main -> INFO 004 Exiting.....
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode query -C $CHANNEL_NAME -n mycc -c '{"Args":["query","a"]}'
2020-10-10 09:19:44.609 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:19:44.609 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
Query Result: 80
2020-10-10 09:19:44.613 UTC [main] main -> INFO 003 Exiting.....
```

Fig. 6. Transaction Execution, a=80 and b=220.

```
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode invoke -o orderer.example.com:7050 --tls SCORE_PEER_TLS_ENABLED --cafile
$ORDERER_CA -C $CHANNEL_NAME -n mycc -c '{"Args":["invoke","a","b","30"]}'
2020-10-10 09:19:58.633 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:19:58.637 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
2020-10-10 09:19:58.669 UTC [chaincodeCmd] chaincodeInvokeOrQuery -> INFO 003 C
haincode invoke successful. result: status:200
2020-10-10 09:19:58.670 UTC [main] main -> INFO 004 Exiting.....
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode query -C $CHANNEL_NAME -n mycc -c '{"Args":["query","a"]}'
2020-10-10 09:20:10.468 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:20:10.469 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
Query Result: 50
2020-10-10 09:20:10.473 UTC [main] main -> INFO 003 Exiting.....
root@ad0eb3690ab8:/opt/gopath/src/github.com/hyperledger/fabric/peer# peer chat
ncode query -C $CHANNEL_NAME -n mycc -c '{"Args":["query","b"]}'
2020-10-10 09:38:19.344 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 001
Using default escv
2020-10-10 09:38:19.344 UTC [chaincodeCmd] checkChaincodeCmdParams -> INFO 002
Using default vscc
Query Result: 250
2020-10-10 09:38:19.351 UTC [main] main -> INFO 003 Exiting.....
```

Fig. 7. Transaction Execution, a=50 and b=250.

IX. PERFORMANCE ANALYSIS AND EVALUATION

To test and evaluate the performance of the Hyperledger Blockchain-enabled IIoT, Hyperledger Caliper version 0.3.2 [33] was used. It is a benchmarking tool to evaluate any Hyperledger Fabric network's performance and provides necessary metrics to analyze the blockchain in terms of success rate, transaction throughput, transaction latency, and resource consumption, including CPU and memory usage. Hyperledger Caliper helps determine scalability, bottlenecks, anomaly, etc. issues in the developed blockchain network.

A. Performance Metrics

1) *Throughput*: Transaction throughput is the rate at which valid transactions are committed and executed successfully. The throughput of the blockchain network is expressed as transactions per second (TPS). The transaction throughput highly depends on the send rate of the network. The send rate is the rate at which the transactions are sent to the network for execution. It is calculated as the following formula (1).

$$Send\ Rate = \frac{(Successful\ Transactions - Failed\ Transactions)}{(Last\ Submitting\ Time - First\ Submitting\ Time)} \quad (1)$$

The Hyperledger Caliper calculates the transaction throughput of the System Under Test (SUT) using the following formula (2).

$$Throughput = \frac{Successful\ Transaction}{(Last\ Submitting\ Time - First\ Submitting\ Time)} \quad (2)$$

The last submitting time is when the transaction gets executed and the first submitting time indicates the time when the transaction was submitted for execution in the network. Throughput only represents successfully committed transactions.

It can be noticed from the above formulae that throughput highly depends on the send rate of the network. If the send rate is high, the transactions throughout will also be high since a high send rate indicates a greater number of transactions to be successfully executed across the network.

2) *Latency*: Latency helps to analyze the amount of time it takes for transactions to be successfully executed or failed if invalid and be effective to be usable across the blockchain network. Hyperledger Caliper calculates the number of committed transactions and when it is successfully executed or failed if invalid. The following formula (3) is used to calculate this metric for each transaction.

$$Latency = Last\ Submitting\ Time - First\ Submitting\ Time \quad (3)$$

Latency indicates the time it takes for both the successfully executed and failed transactions that were invoked. Using this formula, the Hyperledger Caliper shows the maximum, minimum, and average latency of the blockchain network.

3) *Resource consumption*: The Hyperledger Caliper represents the consumption of resources by each peer in terms of the total CPU used in percentage and the amount of memory used to complete their job.

B. Results Analysis

The system analysis was performed with only query transactions as they are generated to communicate with one or more peers of the network, simulating the IIoT device communication. Transaction (Tx) throughput, latency, and resource usage by the peers have been used for the performance metrics. Tables V and VI present the performance analysis results of the Hyperledger Caliper benchmarking tool.

1) *Transaction throughput and latency*: The transaction throughput metric in the Hyperledger Caliper demonstrates the rate of TPS and shows the number of successful transactions. Whereas the transaction latency indicates the time between the submission and execution of a transaction. As shown in Fig. 8, the proposed approach's performance analysis results indicate higher throughput with the increasing number of transactions. It is important to note that the results are taking care of the variation found while running the performance analysis multiple times. As the number of transactions increases, from 24 up to 5000, so is the throughput as Hyperledger Fabric has higher throughput [18]. Though the transactions are faster in a permissioned blockchain than a public blockchain [17], the transaction latency increases with the increasing number of transaction frequencies. If a high-performance server is used, the time required to execute each transaction will further decrease as the signing and encryption will take a shorter time. Moreover, since query transactions do not require consensus from the orderer, but are being handled by the peer itself using Chaincode, the throughput of the query transaction is high, and the transaction delay is comparatively lower than a Public Blockchain [17]. Transaction verification by endorser peer ensures the network's security as tempering the data of the transaction will change the hash of the block causing a smaller number of transactions to be performed and decreasing the throughput of the network. The endorsers can quickly find the tempered block to fail and invalidate. Moreover, since the network allows only a limited number of nodes in the

consensus mechanism, it helps decrease the delay in response and enhances the network's performance.

Some drops and dips in the throughput and latency can be noticed in Fig. 8 graph, as the throughput highly depends on the send rate. If the send rate is high, so is the throughput. But if latency is considered, it has an inversely proportionate relationship with throughput. Therefore, the send rate indirectly affects the latency of the network while directly affecting the throughput. Thus, to produce results with greater accuracy, the send rate has been generated within a range instead of generating with a fixed rate. This allows us to simulate a real IIoT environment where communication will occur with different numbers of transactions and a non-fixed send rate helps to analyze the performance with different scenarios.

2) *Resource utilization*: Hyperledger caliper provides insights on the resource utilization of the blockchain network indicating the usage of CPU and memory. The Docker container of the blockchain network retrieves the container statistics and provides the results of the benchmarking. It is noticeable from Fig. 9 that our system improves the performance of the IIoT network since it utilizes fewer resources. The resources consumed to execute the transactions and device communication are not high. The endorsement peers utilize more resources as they participate in the consensus. Since only a limited number and verified nodes participate in reaching consensus, unlike a public blockchain, our system has low resource usage with high performance.

TABLE V. PERFORMANCE METRICS

Tx Type	Successful Tx	Send Rate (TPS)	Latency (s)			Throughput (TPS)
			Max	Min	Avg	
query	100	7.0	3.50	1.52	2.49	4.4
	400	6.4	3.04	1.12	1.94	5.5
	1300	5.2	2.50	1.03	1.76	4.8
	1600	5.7	3.95	1.01	2.5	5.4
	2500	3.1	1.71	0.8	1.08	2.9
	2800	5.7	2.80	0.90	1.85	4.6
	3700	6.7	3.15	1.10	2.1	5.3
	4000	4.3	1.90	0.85	1.37	3.8
	4900	6.2	2.97	1.01	1.83	4.9

TABLE VI. RESOURCE CONSUMPTION ON AVERAGE

Type	Tx	CPU%(avg)	Memory [MB] (avg)
Docker	100	1.89	76.34
	400	5.32	200.45
	1300	18.67	210.56
	1600	24.33	217.45
	2500	31.23	229.78
	2800	42.55	251.45
	3700	54.55	278.45
	4000	55.89	263.24
	4900	72.33	303.25

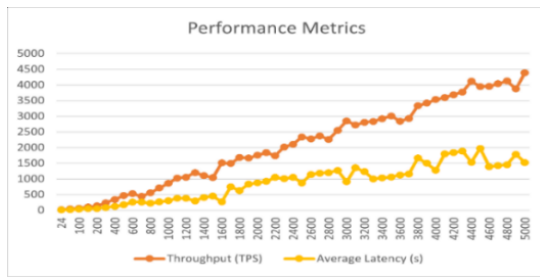


Fig. 8. Performance Metrics.

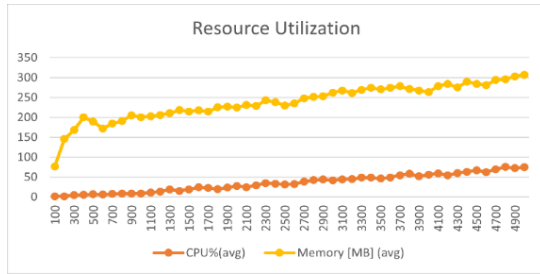


Fig. 9. Resource Utilization.

C. Comparison

The proposed permissioned Hyperledger Fabric Blockchain-enabled IIoT network addresses and improves the security challenges of IIoT as it restricts access and significantly increases throughput, reduces delay in transaction execution, and enhances network performance with required resource usage. There have been studies to use Hyperledger Fabric to improve the security of different industries including [34] where a physical access control management system is developed. Fig. 10 to 15 shows a comparison analysis between using a Hyperledger Fabric for a physical access control device [34] and an IIoT device as our approach. Fig. 10 presents the blockchain network for IIoT devices that have higher throughput than a physical access device [34]. Although [34] performs good with physical control devices, it cannot perform well with IIoT devices in terms of throughput and latency as evident in Fig. 10 and 11. With a lower latency than [34], our approach ensures better performance as it can quickly perform the validation of a transaction. The faster the transaction execution consensus will be received from a limited number of nodes, the more secure the communication medium will be. Therefore, our approach allows low response time with more valid transactions' execution. Fig. 12 presents a combined form of the comparative analysis in terms of throughput and average latency where our approach has a higher throughput and lower latency than [34]. Some of the comparative values of throughput and average latency have been presented in Tables VII and VIII, showing the percentage of increase in throughput and decrease in latency in comparison between IIoT and [34].

However, as presented in Fig. 13 and 14, our approach utilizes more resources than the system of [34]. The authors in [34] have developed an application to control user access through physical devices using Hyperledger Fabric. These physical devices require much less power and space compared to IIoT devices. They are used in certain and specific areas where they are connected to a certain number of other devices.

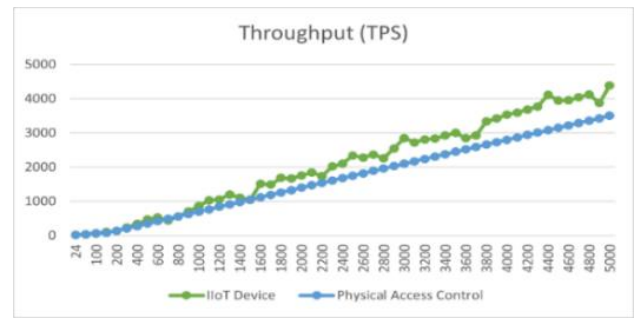


Fig. 10. Comparison of the Performance Metrics: Throughput.

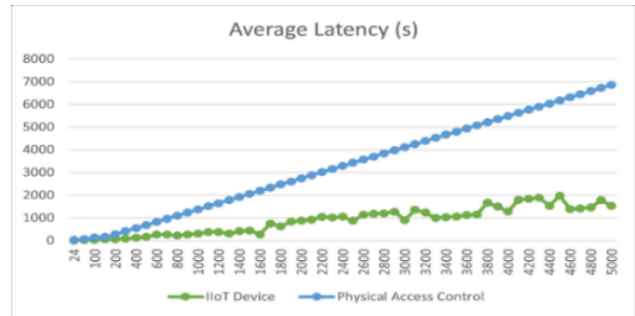


Fig. 11. Comparison of the Performance Metrics: Average Latency.

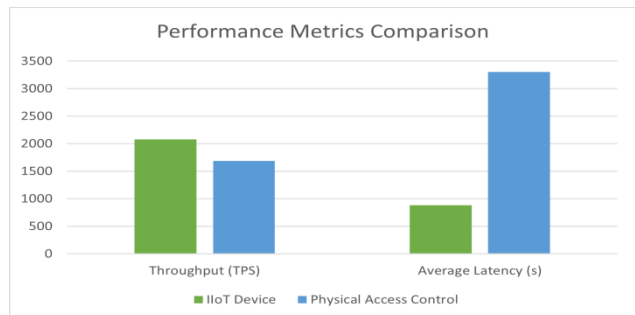


Fig. 12. Comparison of the Performance Metrics.

On the other hand, numerous IIoT devices are used in huge industrial sectors, sharing information and requiring real-time data processing. They need specific storage capability and huge processing power to be able to communicate efficiently [35]. A simple physical access control device merely logs the user access records. In contrast, IIoT devices not only share information but also produce data and provide improved business decision insights, requiring them to comprise with high processing power and memory usage [35]. Therefore, to meet the high performance requirements of IIoT devices [35], our approach uses more resources and secures the network from malicious attackers. It also allows using any consensus algorithm as per IIoT requirement. The algorithm ensures the participation of only a limited number of nodes in reaching a consensus. Consequently, it assures low response time and required usage of resources that helps in having an IIoT permissioned blockchain network with high performance and security. The comparative analysis of the resource utilization by our approach and [34] is highlighted in Fig. 15.

TABLE VII. THROUGHPUT COMPARISON

Throughput (TPS)			
Tx	IIoT/Our Approach	Physical Access Control [34]	Performance
100	62.85	70	10% Decrease in Throughput
400	343.75	280	23% Increase in Throughput
1300	1200	910	32% Increase in Throughput
1600	1515.789	1120	35% Increase in Throughput
2500	2338.71	1750	34% Increase in Throughput
2800	2259.649	1960	15% Increase in Throughput
3700	2926.866	2590	13% Increase in Throughput
4000	3534.884	2800	26% Increase in Throughput
4900	3872.581	3430	13% Increase in Throughput

TABLE VIII. AVERAGE LATENCY COMPARISON

Average Latency (s)			
Tx	IIoT/Our Approach	Physical Access Control [34]	Percent Improvement
100	35.57	137.2	74%
400	121.25	548.8	78%
1300	300	1783.6	83%
1600	267.89	2195.2	88%
2500	870.96	3430	75%
2800	1200	3841.6	69%
3700	1159.70	5076.4	77%
4000	1274.42	5488	77%
4900	1785.65	6722.8	73%

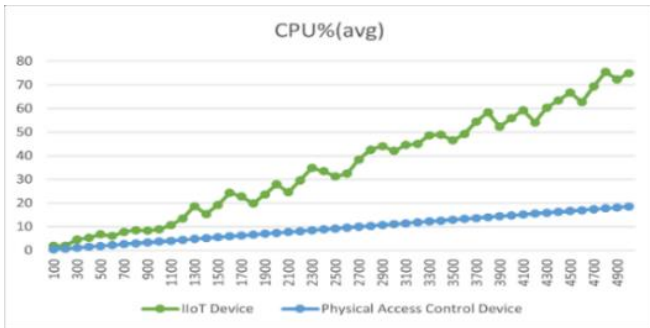


Fig. 13. Comparison of CPU usage.

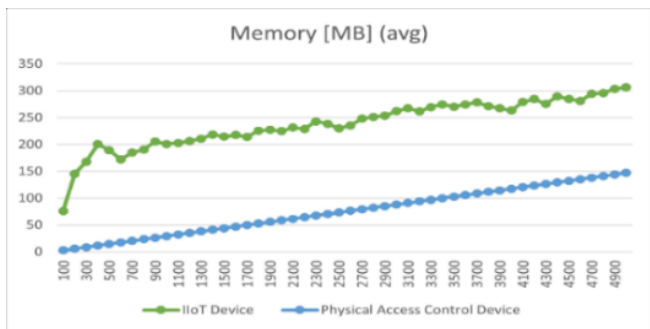


Fig. 14. Comparison of Memory usage.

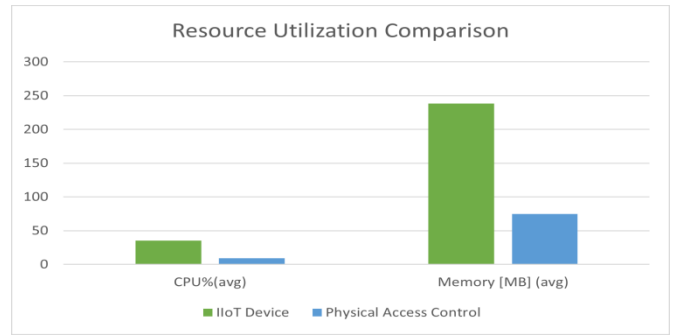


Fig. 15. Comparison of Resource usage.

X. CONCLUSION

With the emerging and diversified use of IIoT, it is important to address the security vulnerabilities of this technology. The integration of blockchain and IIoT can play a vital role in overcoming the IIoT security limitations. Therefore, this paper proposes to use a Hyperledger Fabric Blockchain to secure the IIoT device communication and ensure data is stored, accessed, and monitored by only authorized parties. The Hyperledger Fabric Blockchain-enabled IIoT uses a CA to issue certificates to the identities and authorize them to perform transactions, and MSP defines user access roles based on the certificates. The peers of the blockchain network validate transactions using the definition provided in the Chaincode and generate the transactions as blocks in the network if verified by following Chaincode, and prevent the double-spending attack. It also updates the rest of the peers and provides consistency across the whole network. This paper implements the proposed solution and performs an extensive evaluation in terms of throughput, latency, and resource utilization, to analyze the security of the communication medium. Using the optimum values, the performance analysis indicates that the Hyperledger Fabric blockchain is suitable for the IIoT network that improves the security and ensures only authorized and authenticated identities are participating in device communication.

XI. FUTURE WORK

This study focused on securing the IIoT device communication medium using Hyperledger Fabric and ensuring that security management remains intact. In future research, an extensive study will be performed in guaranteeing the proposal follows the CIA triad and is available to only authorized users. The need for such future work is required to solve and improve the utilization of resources by the IIoT devices. It will also include the study and future analysis on the security vulnerabilities of blockchain that might affect the IIoT environment. Furthermore, future studies will be dedicated to discovering any exploitable bugs in the proposed network.

REFERENCES

- [1] S. Yeasmin, A. Baig. "Permissioned Blockchain-based Security for IIoT." In 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), pp. 1-7. IEEE, 2020.
- [2] A. Esfahani, G. Mantas, R. Matischek, F.B. Saghezchi, J. Rodriguez, A. Bicaku, S. Maksuti, M.G. Tauber, C. Schmittner, J. Bastos. "A Lightweight Authentication Mechanism for M2M Communications in Industrial IoT Environment". IEEE Internet of Things Journal, vol. 6, no. 1, pp. 288-296, Feb. 2019.

- [3] W. Yang, S. Wang, X. Huang, Y. Mu. "On the Security of an Efficient and Robust Certificateless Signature Scheme for IIoT Environments". *IEEE Access*, vol. 7, pp. 91074-91079, 2019.
- [4] Y. Zhang, R. H. Deng, D. Zheng, J. Li, P. Wu and J. Cao, "Efficient and Robust Certificateless Signature for Data Crowdsensing in Cloud-Assisted Industrial IoT," in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5099-5108, Sept. 2019, doi: 10.1109/TII.2019.2894108.
- [5] J. Huang, L. Kong, G. Chen, M. Wu, X. Liu, P. Zeng. "Towards Secure Industrial IoT: Blockchain System With Credit-Based Consensus Mechanism". *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3680-3689, June 2019.
- [6] W. Liang, M. Tang, J. Long, X. Peng, J. Xu, K. Li. "A Secure FaBric Blockchain-Based Data Transmission Technique for Industrial Internet-of-Things". *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3582-3592, June 2019.
- [7] Y. Lu, X. Huang, Y. Dai, S. Maharjan, Y. Zhang. "Blockchain and Federated Learning for Privacy-Preserved Data Sharing in Industrial IoT". *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4177-4186, June 2020.
- [8] Q. Wen, Y. Gao, Z. Chen, D. Wu. "A Blockchain-based Data Sharing Scheme in The Supply Chain by IIoT". In 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), Taipei, Taiwan, 2019, pp. 695-700.
- [9] J. Wan, J. Li, M. Imran, D. Li, Fazal-e-Amin. "A Blockchain-Based Solution for Enhancing Security and Privacy in Smart Factory". *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3652-3660, June 2019, doi: 10.1109/TII.2019.2894573.
- [10] T. Alladi, V. Chamola, R. M. Parizi, K. R. Choo. "Blockchain Applications for Industry 4.0 and Industrial IoT: A Review". *IEEE Access*, vol. 7, pp. 176935-176951, 2019.
- [11] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, M. Gidlund. "Industrial Internet of Things: Challenges, Opportunities, and Directions." *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724-4734, Nov. 2018.
- [12] P. Sethi, S. R. Sarangi. "Internet of things: architectures, protocols, and applications". *Journal of Electrical and Computer Engineering*, 2017.
- [13] W. Liang, M. Tang, J. Long, X. Peng, J. Xu, K. Li. "A Secure FaBric Blockchain-Based Data Transmission Technique for Industrial Internet-of-Things". *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3582-3592, June 2019, doi: 10.1109/TII.2019.2907092.
- [14] S. Yeasmin, A. Baig. "Unlocking the Potential of Blockchain". In 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 2019, pp. 1-5.
- [15] I. Makhdoom, M. Abolhasan, W. Ni. "Blockchain for IoT: The challenges and a way forward". In ICETE 2018-Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, 2018.
- [16] A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram. "LSB: A Lightweight Scalable Blockchain for IoT security and anonymity." *Journal of Parallel and Distributed Computing* 134 (2019): 180-197.
- [17] N. Andola, M. Gogoi, S. Venkatesan, S. Verma. "Vulnerabilities on hyperledger fabric," *Pervasive and Mobile Computing*, 59, 101050, 2019.
- [18] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, S. Muralidharan. "Hyperledger fabric: A distributed operating system for permissioned blockchains". In *Proceedings of the thirteenth EuroSys conference*, pp. 1-15. 2018.
- [19] Corda, Available online: <https://docs.corda.net/key-concepts-notaries.html>.
- [20] A. Vyas, L. Nadkar, S. Shah. "Critical Connection of Blockchain Development Platforms". In *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2019, ISSN: 2278-3075, Volume-8, Issue- 9S2.
- [21] T. Koens, S. King, M. van den Bos, C. van Wijk, A. Koren, "Solutions for the Corda Security and Privacy Trade-off: Having Your Cake and Eating It".
- [22] R3 Corda, Available online: <https://www.r3.com/corda-platform/>.
- [23] M. Hearn. "Corda: A distributed ledger". *Corda Technical White Paper*, 2016.
- [24] Openchain, Available online: <https://www.openchain.org/>.
- [25] T. T. A. Dinh, R. Liu, M. Zhang, G. Chen, B. C. Ooi J. Wang. "Untangling Blockchain: A Data Processing View of Blockchain Systems". In *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1366-1385, 1 July 2018.
- [26] D. Schwartz, N. Youngs, A. Britto. "The ripple protocol consensus algorithm", *Ripple Labs Inc White Paper*, 2014, 5(8).
- [27] Ripple, Available online: <https://ripple.com/xrp/>.
- [28] P. Sajana, M. Sindhu, M. Sethumadhavan. "On blockchain applications: hyperledger fabric and Ethereum". *International Journal of Pure and Applied Mathematics*, 118(18), 2965-2970, 2018.
- [29] P. Thakkar, S. Nathan, B. Viswanathan. "Performance Benchmarking and Optimizing Hyperledger Fabric Blockchain Platform". In 2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Milwaukee, WI, 2018, pp. 264-276.
- [30] Hyperledger Fabric, Hyperledger Fabric, Available online: <https://hyperledger-fabric.readthedocs.io/en/release-2.0> [Accessed on 01 March 2020].
- [31] Hyperledger Fabric, 2018, Available online: <https://cloud.ibm.com/docs/blockchain?topic=blockchain-hyperledger-fabric> [Accessed on 01 March 2020].
- [32] Sadeghi, C. Wachsmann, M. Waidner. "Security and privacy challenges in industrial Internet of Things". In 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, CA, 2015, pp. 1-6.
- [33] Hyperledger Caliper, Hyperledger Caliper Documentation, Available online: <https://github.com/hyperledger/caliper> [Accessed on 1 November 2020].
- [34] S. Rouhani, V. Pourheidari, R. Deters. "Physical Access Control Management System Based on Permissioned Blockchain". In 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Halifax, NS, Canada, 2018, pp. 1078-1083.
- [35] G. Caiza, M. Saeteros, W. Oñate, M. V. Garcia. "Fog computing at industrial level, architecture, latency, energy, and security: A review". *Heliyon*, 2020, 6(4), e03706.

A Hybrid Technique based on RSA and Data Hiding for Securing Handwritten Signature

Yaser Maher Wazery¹, Shimaa Gamal Haridy², AbdElmegeid Amin Ali³
Faculty of Computers and Information, Minia University

Abstract—Data exchange has been significantly encouraged by the development of communication technology and the wide use of social media over the Internet. Therefore, it is important to hide the data transmitted, especially the data that requires a person's signature. Where the signature is an increasingly needed item that is used in our daily life to achieve some paper-based authentication in departments, the individual himself needs the signature. Cryptography and steganography are commonly considered to be the most important data hiding methodologies. Steganography is used to hide the secret message in the carrier media, such as text, audio, video, and image files, without the carrier media being distorted, and cryptography is used to conceal the purpose of the secret message. A hybrid data hiding (image steganography) and encryption technique is implemented in this research on the time domain. The secret handwritten signature image is first encrypted using the public key algorithm (RSA) in the proposed technique, then randomly inserted using Embedding data process to be concealed in one of the last three bits of that pixel (1st Least Significant Bit, 2nd LSB, and 3rd LSB) based on mathematical randomized formula over all pixels of the carrier media (image). It is assumed that the process of randomization will increase the protection provided by the technique. The suggested technique is implemented on gray level cover images. As a consequence of the random scattering of bits and using encryption, it is noted that the proposed technique achieves enhanced data hiding results in terms of performance, protection, and imperceptibility properties and the histogram of the proposed technique is better and provides more protection and security than the ordinary sequential Least Significant Bit (LSB).

Keywords—Image Steganography; LSB; Data Hiding; Security; Embedding data; Cryptography; RSA; Handwritten signature

I. INTRODUCTION

With the exponential growth of technology, digital communication and social media, data protection has become very critical. In data communication, security problems during transmission are required to deal with it. The specifications of secure communications are therefore important. Reliable personal identification/authentication is important because of the increasing importance of security technologies. The need for safety and access restrictions is important to safeguard the data transmitted, especially data that requires person's signature. Where the signature is an increasingly needed item used in our daily life to achieve some paper based authentication in departments. So in order to protect information over communication networks, there are two common types of techniques: cryptography and hiding information [1], which typically complement each other. The term hiding can either make the information undetectable (as in watermarking) or keep the information hidden (as in steganography). On the other hand, cryptography [2] is a method used to maintain the

confidentiality of the content of the message. For the purpose of encrypting and decrypting sensitive data, several different approaches have been suggested and implemented.

- **Cryptography has two classification types:**

- Classical cryptography: In this category the letters of the original message are encoded using either substitution techniques (each letter is replaced with another letter depending on key) or transposition techniques (reorder the letters of the original message to obtain the cipher text).
- Modern cryptography: There are two types of algorithms in this category, (symmetric and asymmetric encryption).

Symmetric Encryption: The sender and the receiver must have the key to encrypt or decrypt message respectively. The symmetric encryption [3] has one key in the two sides for the message, so the sender conceals the message using a shared key with the receiver and the receiver decrypt the message using the same shared key.

Asymmetric Encryption (Public Key Cryptography): This approach has two different keys (private and public) used for the message encryption and decryption. Anyone can know the public key, and it can be used to encrypt messages and check signatures. Although the receiver is only aware of the private key, it is used to decrypt messages and sign signatures [4], as seen in figure 1.

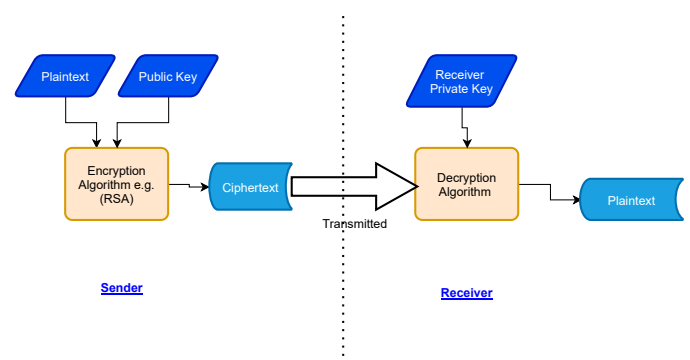


Fig. 1. Public Key Cryptosystem.

The use of public key cryptosystem like RSA provides the following advantages:

- The encryption key is public and distinct from the decryption key in a public-key cryptosystem, which

is kept secret (private) but both are mathematically related.

- An RSA user generates and publishes, along with a supplementary value, a public key based on two large prime numbers. The prime numbers are secretly stored.
- Via the public key, messages can be encrypted by anyone, but can only be decoded by someone who knows prime numbers and private key.
- RSA's protection relies on the practical complexity of factoring two large prime numbers (factoring problem) into the product and modular arithmetic.

Steganography uses a carrier medium such as text, video, image, and audio file to cover the hidden data, according to figure 2. It is must for steganography to have some message to be embedded [5] and a cover medium in which the embedded message is hidden.

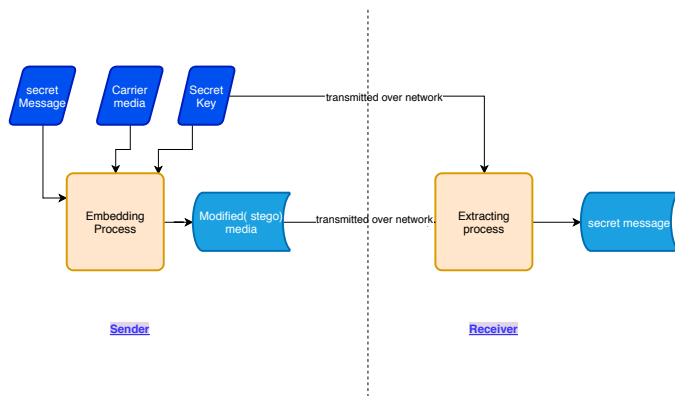


Fig. 2. Steganography Model.

The system of steganography should satisfy imperceptibility, high capacity and security[6], the key factors that influence steganography and its usefulness are these three aims. The accuracy of the image is also a significant objective for steganography. Two widely recognized methods are available to measure image quality: the Peak Signal-to-Noise Ratio (PSNR) and the Mean Squared Error (MSE) indicators. In such a case, PSNR is generally used to precisely detect the degree of corruption in the changed images compared to the carrier media. Whereas MSE is the solution to explain the distinction between two distinct images. Equations 1 and 2 [7] define MSE and PSNR.

$$MSE = \left(\frac{1}{S}\right) \sum_{i=1}^S (Z_i - Z'_i)^2 \quad (1)$$

$$PSNR = 10 \log_{10} \frac{I^2}{MSE} \quad (2)$$

Where Z_i is the index of the i^{th} cell pixel in the carrier image, Z'_i is the index of the i^{th} cell pixel in the modified image, where the parameters S is meant to be the size of the both images and I is the upper bound of the pixel value, for gray level images (8-bits per pixel), $I = 255$.

The protection obtained by steganography to mask sensitive data inside a cover media depends on the presumption that no one may assume any secret data is available. However, if someone discovers a difference in the cover media [8], it is possible to discover sensitive data. Therefore, before concealing it in the cover media, it is preferred to use another approach such as cryptography to encrypt the sensitive data, this would ensure that even though the embedded data is found [9], no one will know its meaning. Therefore, we should take advantage of hybridization the two strategies for better protection. In general, steganography is the science of hiding information through a certain process in another type of cover media, i.e. text, video, audio and image[2]. While cryptography is the technique that use mathematics to convert intelligible data into unintelligible form to keep messages safe. The main description of steganography process is shown in figure 1. This study focuses on Cryptography and Steganography based on images where the embedding is done with the encrypted key concealed in the Object cover after encrypting the message. As the other layer of this encryption scheme, cryptography primarily encrypts the hidden plain text / image, converting it to cipher text/ image.

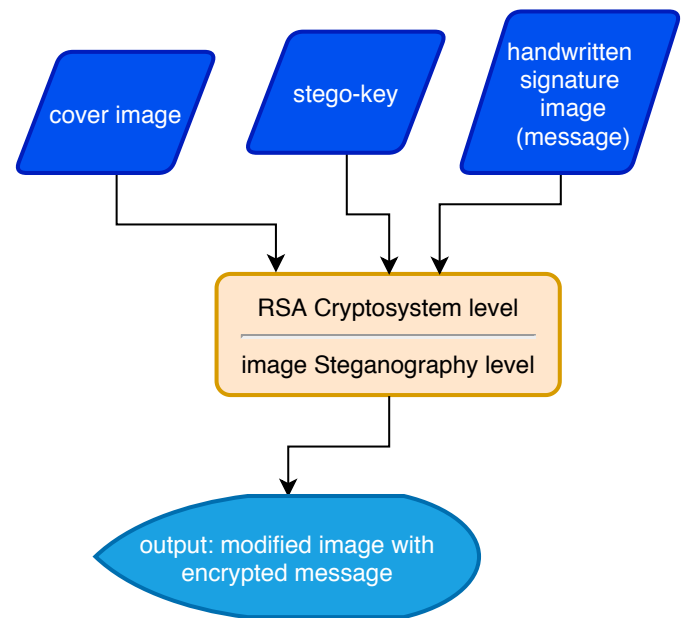


Fig. 3. Proposed Model.

This study introduces a 2-level protection framework for crypto-stego using steganography of the image base as a dependent level and cryptography of RSA as an independent assurance level. The sensitive data passes through the crypto level in the proposed security strategy, followed by the steganography level, which results in the output file as Stegoimage. The main description of the process using these two level techniques is shown in Figure 3. In this paper a security technique based on steganography and public key (RSA) encryption as two levels is proposed. The cover media is a gray level image. The cover image is hiding the secret ciphered message (handwritten signature) by RSA using a modified scattered Least Significant Bit (LSB). The most common types of images used in our daily life situations for handling paper based work is gray level imaging after scanning because it is cheaper and available in

most of places.

II. LITERATURE WORK / REVIEW

In [10], by using scattered LSB, the authors implemented a technique that hides a fingerprint image as (message) in a face image as (covered image). They first use a password to encrypt the fingerprint image bits, then embed those bits using LSB but not linearly, rather a pseudo random number generator (PRNG)[11] that depends on the password is used to pick the position of the pixel in which the bit would be inserted. By using this structure, without knowing the password, it is not possible for anyone to try reading the secret data to get the hidden version (not even the encrypted version). The observations of this proposed technique explained that scattered LSB introduces a large value of PSNR comparing to sequential LSB, which means image quality of scattered LSB is higher than of the values obtained by sequential LSB.

In [7], they presented a method that arbitrarily (depending on the mathematical equation) embeds the character of the secret message into the carrier media in just three pixels. A colored picture reflecting three levels (Red, Green, and Blue) is the carrier media. During the first iteration, only two tiers (Green and Blue) are used in the embedding operation. In the second iteration, only the blue tier should be included. Two more stages (Green and Blue) and one stage (blue tier) are used in the following iterations and so on. The use of this approach leads to protection improvements and better performance for PSNR values. The secret information embedding is performed randomly. The pixel position decision is made by means of a PRNG [12] rather than ordinary linear matter. The message bits must be hidden in a shape of (3-2-3) during the insertion process. Where the first 3-bit of the original message is entered at the first random position of the pixel ((2-bit of the blue level at (7th bit) and bit at (8th bit)), (and the third bit of the 3-bit of the green level is inserted at the 8th position of the message). Then, in the place of the second random pixel (7th bit) and (8th bit) of the Blue level, the successor 2-bits obtained from the original message (4th and 5th bits) are entered. After that, in the third random pixel location ((2-bits in the blue level at (7th bit) and (8th bit)), (1-bit at (8th bit) of the green level), and so on, the final 3-bits obtained from the original message (6th, 7th and 8th bits) should be added. The outcomes of the proposed method explains that it incorporates greater PSNR value and greater ability for embedding than sequential LSB.

A hybrid data hiding (HDH) technique applied to the medical imaging field was suggested by the authors in [13]. HDH combines Hamming (3, 2) + 1 and original LSB techniques with the Optimal Pixel Adjustment Process (OPAP) system used to encode patients' secret information. HDH strengthens the process of 'Hamming + 1'. Moreover, the output and the capability of changing The photos have been enhanced. Comparing to other similar techniques, the results of the implemented approach showed an enhancement in the hiding ability carried out by this technique. In addition, the accuracy of the updated images remained greater than 50 dB in the medical sector for the proposed scheme images.

In [14], the authors suggested new and creative audio steganography for the purpose of popularizing the use and products of IoT services. The proposed solution provides

a more stable IoT climate. This study focuses on audio steganography, concentrating on the hidden message followed by the adoption in the originally provided audio stream of the optimized audio embedding technique (OAET) for shuffling bit embedding replacement. for concealing the hidden message, a random bit selection is applied by the technique used in[14] to conceal the necessary data in the farthest part of the audio stream. Their suggested technique showed improvements in the robustness of the inserted audio stream, and the outcomes showed a decrease in the distortion impact. The process uses WAV as the default format for the original stream of audio. The results of this paper also showed that the quality of the audio file is better than the standard LSB after implementing the proposed technique, and provides high-level protection.

In [15], the authors suggested an improved method to protect confidential personal computer text data that benefits from the combination of(cryptography and steganography). The protection of the system is provided by the inclusion of RSA cryptography followed by video based steganography as two sequential layers to ensure the best possible safety. The implementation of the framework starts with the user entering a secret confidential text data message and a secret key for the cryptography level The software transforms each character of the confidential secret text within this level method into an array of binary to be encrypted using RSA. The second level, i.e. the steganography level, also demands the cover media for an RGB video frame, so that its pixels are also transformed into binary form. There are 3 channels of any pixel in the RGB video frame (red, green and blue) displaying a byte of each. The authors used 3 bits of hidden data to be embedded in each pixel using the least significant bits (LSB) of video-based steganography in their paper. In order to explore the relationship between protection, capability and data dependence, the study modeled the system and implemented it for testing. The experiments involved testing of data protection in 15 different video sizes that yielded interesting results in comparison with the existing method in [16]. this study reinforced capability vs. security, as an inevitable trade off was implemented. The tests included all possibilities for using number of bits to be concealed in one pixel (1-LSB, 2-LSB, and 3-LSB) security acceptance methods describing their impact on the cover video. The major result proved applicable to the implementation of the 3-LSB approach to provide acceptable protection with realistic capacity preferred between 1-LSB and 2-LSB methods.

In [17], the authors suggested a hybrid technique to secure the secret data that use the behaviors of steganography (LSB, raster scan technique) and cryptography (symmetric key). Using a symmetric key cryptography technique, which is content-based and uses the block cypher principles, the secret text will first be encrypted, but the size of the block in this technique is not determined, depending on the length of the term (word). Secondly, the embedding process for ciphered data in RGB carrier image at the 3 planes R, G and B is occurred as follows: Using modified LSB replacement by XORing hidden data bits with cover pixel bits, 2 bits will be embedded in 2 LSB red plane, then 2 bits will be inserted using raster scan technique in 2 LSB of green plane (hide from left to right in the first scan and right to left in the next scan and so on) by XORing hidden data bits with covering pixels, then using raster scanning technique (hide from top to bottom in the

first scan and from bottom to top in the next scan and so on) 4 bits will be inserted in 4 LSB of blue plane by XORing hidden data bits with bits of cover pixels. This process is repeated until the entire text of the cipher is concealed in the carrier image. The statistical results provide that no observation difference between the cover image and the Stego image. In analysis method, MSE and PSNR parameters are calculated and correlated with the performance of current technique [18], and the results show that the proposed work has variability in the PSNR and the degree of protection is also very high.

In [19], the authors suggested a hybrid technique to secure the image and text using the combination of cryptography and steganography (RSA, LSB and DWT). In this research, a gray level image is taken as a carrier image, then the replacement of the LSB bits on the cover image is applied after choosing the cover image. Using the RSA encryption method, the secret data will be translated into cipher text behind the encrypted image, the encrypted text is concealed. So The original message is protected by two layers of security. Firstly, the secret message itself is encrypted, and then the cover image is encrypted as well. It's then inserted in the original image. The LSB extraction method is used in the decoding process to get the message bits. The bits are then removed from the location in the same order as they were embedded, when the position of the bits has been defined. Extract encrypted images from the DWT cover image and decrypt text with the private key of the recipient using the RSA technique. The results show that MSE of images used is less by comparing the MSE values of all images, and PSNR of images is higher than the present technique [17] in the proposed technique. The result of images derived from entropy indicates that the entropy of the modified image is relatively higher than the cover image. This is due to the inclusion of more secret details to the cover image.

In [20], the authors suggested a scheme to produce a reliable and stable message transfer technique such that private and confidential information can be transmitted over the network in a secure manner. The proposed method is applicable to gray scale images. The pixels' 7^{th} bit is subjected to a mathematical function. The 7^{th} bits of the selected pixel and the pixel +1 value are obtained, and 2 bits of the message can be extracted from each pixel using a combination of these two values. 00, 01, 10, and 11 are the four possible variations. This approach has many benefits, including two bits of message are stored in each pixel and the technique's independence from the 8^{th} digit. When inserting the data into the image file, the pixel value will shift by a maximum of +2 and -2. The results when compared to other approaches, show a high PSNR and a low MSE.

III. THE PROPOSED TECHNIQUE

The proposed technique for securing and hiding the handwritten signature image is illustrated and clarified in this section.

The proposed scheme's goal is to create a safe and robust message transfer technique such that private and sensitive information (handwritten signature) can be transmitted over the network in a secure manner without being vulnerable to unintended recipients and attacks. The proposed method is applicable to gray scale images. We introduced a multilevel

security paradigm first by securing through RSA then providing a sophisticated embedding by applying a randomized positioning for choosing cover media bit's position and value. The proposed technique can be described as two stages; the first one is encrypting the handwritten then embedding / inserting and the second is extraction and restoring the handwritten signature image. This technique aids in overcoming steganography's weaknesses in traditional LSB to a greater degree. Embedding Algorithm & Extraction Algorithm are shown in figures 4, 5 respectively.

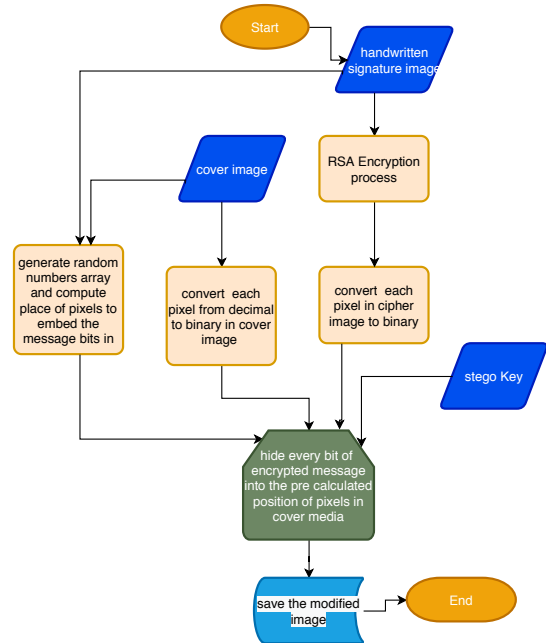


Fig. 4. Encryption and Embedding the Handwritten Signature Image Process.

A. Encrypting and Embedding Stage

The first algorithm will be used to conceal the secret handwritten signature as image randomly (depending on a mathematical equation) in the (1^{st} LSB, 2^{nd} LSB and 3^{rd} LSB) over the carrier media which is a gray level image after encrypting using RSA encryption Algorithm. The inputs of the insertion algorithm are the handwritten signature's image and the carrier media; the first technique is divided into two parts (Encryption using RSA and embedding using modified LSB):

1) *RSA Encryption*: The RSA algorithm is a method of public key encryption and is known as the most secure form of encryption. It was invented in 1978 by Rivest, Shamir and Adleman, the RSA algorithm has the name of them. The steps of encrypting the handwritten signature image are shown in figure 6 and the algorithm is shown in (algo: 1).

Algorithm 1: RSA Encryption

Input: Handwritten signature image;
Output: Ciphred signature array;
 First, the two key pairs (public and private) are generated by:

Selecting two different prime numbers (a and b).
 Calculate the product for public and private keys (n) by the equation [21]:

$$n = a * b \quad (3)$$

Calculate the totient ($\phi(n)$) [21]:

$$\phi(n) = (a - 1) * (b - 1) \quad (4)$$

Select an integer value e for public Key, such that $1 < e < \phi(n)$ and (e , is relatively prime to $\phi(n)$) if they share no common factors other than 1; $gcd(e, \phi(n)) = 1$.
 Calculate the private key d to fulfill the congruence relation $e.d \equiv 1 \pmod{\phi(n)}$.

The modulus n and the encryption exponent e generate the public key.

a, b and the private exponent d generate the private key which must be kept secret.

The handwritten signature image is converted to one dimension array of decimal numbers.

then is encrypted using RSA public key by RSA encryption equation [21]:

for $i = 1 \leq message_Length$ **do**

$$C_{handwritten}(i) = (M_{handwritten}(i))^e \pmod{n}; \quad (5)$$

end for

where $C_{handwritten}$ is the handwritten signature after applying RSA encryption and $M_{handwritten}$ is the main handwritten signature.

return ciphred handwritten array;

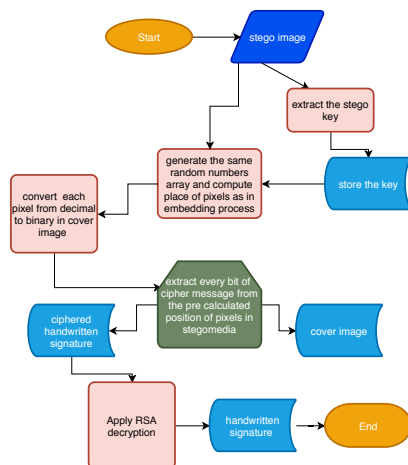


Fig. 5. Extraction and Decryption the Handwritten Image Process.

2) *Embedding using Modified LSB:* After making RSA encryption the cipher handwritten signature are entered to the embedding algorithm with the 16-bit carrier cover image and a secret key, the algorithm is stepped in (algo: 2) .

Algorithm 2: Embedding using Proposed LSB

Input: Ciphred handwritten image, Cover image, Stego-key;

Output: Stego image;

Each pixel of ciphred handwritten signature's is converted from decimal to binary of 16-bit length generating $binary_Image$.

Compute place of pixels n_Blocks to embed the message bits in (this random array is generated as the same in the destination side).

$$n_Blocks = \frac{image_Length}{s_length} \quad (6)$$

Where $image_length$ is the size of cover image and s_length is the size of handwritten image.

Generate and calculate the random array R that ranges from 6 to 8 the last three bits of LSB

$$R = \sum_{i=1}^{s_Length} rand(6 : 8) \quad (7)$$

The same R are generated at the extracting algorithm by using this matlab function: $rng(seed, generator)$ where rng is a Control function that handles random number generation, Seed is the parameter used to seed the random number generator using a non-negative integer value, and a generator additionally specify the type of the random number generator.

Then the pixels of carrier media is transformed to its ASCII binary.

Calculate N the length of the ciphred signature.

Convert the cover image matrix to column.

Initialize $k = 1, t = 1$;

for $i \leq image_Length$ **do**

if $k \leq N$ **then**

Every bit of the handwritten signature's image is hidden after a calculated number of pixels' blocks from equation 6 at the last three bits LSB in the pixel of cover image randomly based on the equation 7.

$Stego_Image =$

$LSB(R(t), cover_image(i), binary_image(k));$

Increment $k = k + 1$;

Increment $t = t + 1$;

end if

Increment $i = i + n_Blocks$;

end for

convert the column $Stego_Image$ to matrix;

After hiding the signature, a secret key is hidden into the modified image $Stego_Image$ that is known for both sender and receiver to be able to extract the handwritten signature image after receiving.

write the new image($Stego_Image$)

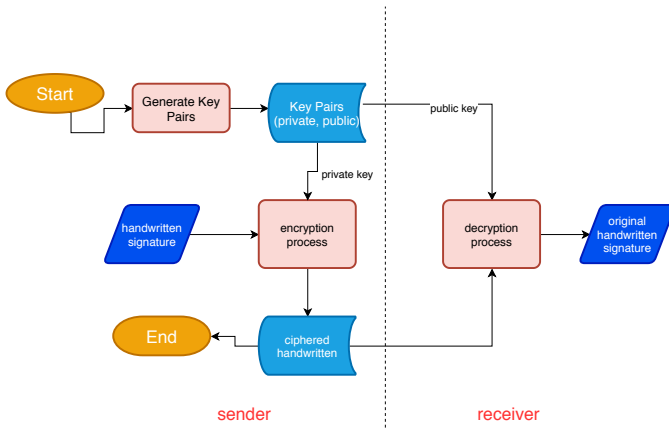


Fig. 6. RSA Model.

B. Extracting and Decrypting Stage

The second algorithm will be used to extract the ciphered secret handwritten signature from a stego-image using the same random numbers those are generated in the embedding process (depending on a mathematical equation) over the modified gray level image and then decrypting it using RSA decryption Algorithm. The second stage is divided into two parts (extracting using modified LSB and decrypting using RSA which are shown in algo: 3 and algo: 4): The inputs to the extraction algorithm are (the *stegokey* and the *Stego-image*) for extraction and (private key of RSA (d, n)) for decryption.

1) Algorithm of Extraction part:

2) Algorithm of Decryption part:

IV. EXPERIMENTAL RESULTS

The use of RSA cryptosystem in the proposed technique is according to:

- A key exchange is not necessary in asymmetric RSA and this improves the security of the algorithm.
- For the cryptographic process, RSA uses factorization that significantly decreases the algorithm's speed.
- Much faster than RSA are symmetric algorithms but a protected encryption scheme (Cryptosystem as RSA) is needed to protect against brute force attacks and differential plaintext-cyphertext attacks. The handwritten signature is important to be secret during transmission so the RSA is used.

The covered images used are 8-bit gray scale images. The handwritten signature images to be encrypted and hidden are taken from a data-set in¹ kaggle website[22]. The cover image contains information about the hidden file, such as hidden file size. While developing the technique the concern was about four different sizes of cover images (papers of size A4, A3, A5) and a template cover images are used such as (Lena, Pepper and Pears) of size 512*512, and fixed size for handwritten signature image (40*80) are used of type uint16 to achieve more security after using RSA for bigger values. After

¹<https://www.kaggle.com/divyanshrai/handwritten-signatures>

Algorithm 3: The Extraction algorithm

Input: *Stego-image* and *Stego-key*;

Output: ciphered handwritten image as column; exclude the message size from the *Stego-image* depending on *Stego-key*.

convert the image to one array (column).

generate same random array to retrieve the bits and pixels places of the embedding message (1st, 2nd or 3rd LSB) to calculate *RG* (Random array of bits (6:8) from equation (7)) and *noBlocks* (no. of blocks representing the location of pixels that the message was embedded in, from equation (6)).

$k = 1, t = 1;$

for $i \leq imageLength$ **do**

if $k(ismessageindex) \leq messageLength$ **then**

$StegoPixelBits = \text{convert } stegoImage \text{ from decimal to binary of 8 bit length};$

$secretBits(k) = StegoPixelBits(RG(t))$

increment $k = k + 1;$

increment $t = t + 1;$

end if

increment $i = i + noBlocks;$

end for

combine each 16 bit and add it in one pixel in *secretImage* array.

define $tt = 1;$

define $index = 1;$

for $i = 1 \leq messageLength$ **do**

for $j = 1 \leq 16$ **do**

$SecretImage(index) += secretBits(i);$

$j = j + 1;$

end for

$index = index + 1;$

increment $i = i + 16;$

end for

convert the secret image array *SecretImage* from binary to decimal.

Algorithm 4: RSA decryption

Input: Ciphered handwritten image array;

Output: Original handwritten image;

Then applying the RSA decryption using the private key (d, n).

for $i = 1 \leq messagelength$ **do**

Calculate message by the equation of decrypting RSA [21]:

$$M_{handwritten} = C_{handwritten}^d \text{ mod } n \quad (8)$$

$Calculatemessage(i) =$

$power(cipheredSecretImage, d) \text{ mod } n;$

end for

Convert the column array (message) to image (the original handwritten signature image).

encrypting the handwritten signature using RSA, its bits does not use sequential LSB but it is distributed randomly according to the use of a PRNG that depends on the carrier media size and the handwritten signature image size. This PRNG determine two things: the placements of pixels in covered image to insert the handwritten image's bits and the bit number in that pixel to do scattered LSB.

A. PSNR value:

PSNR is defined as the ratio between the desired signal power and the noise signal power (signal that corrupts the main signal). A higher PSNR value shows that the image has better quality. PSNR value of the proposed technique was calculated for original and modified images and results are clarified in table I. While experimenting the proposed algorithm it was

TABLE I. PSNR AND MSE OF THE PROPOSED TECHNIQUE IN DIFFERENT SIZES OF COVERED IMAGE AND SAME SIZE OF HANDWRITTEN IMAGE

Cover image template size	Cover image size	Handwritten signature image size	PSNR	MSE
A4	3508*2480	40*80*16 bit depth	64.8506	0.0213
A3	4961*3508	40*80*16 bit depth	67.8931	0.0106
A5	2480*1748	40*80*16 bit depth	61.9136	0.0419

vital to compare its results to the ordinary LSB(1st LSB) and the technique in [17]. The proposed technique is dealing with gray cover image which has less pixels than a color cover to embed the secret data, also it is dealing with last three bits of pixel (1st, 2nd and 3rd LSB) randomly which has better security than sequential LSB (1st LSB) so in the case of discovering the hiding message, the extraction of the proposed technique will be hard than using 1st LSB and if it is extracted the message is encrypted using RSA. The MSE and PSNR value of the technique [17] was provided in comparison to the proposed and the 1stLSB in tables II and III respectively. It is normal that PSNR of 1st LSB has more value than any technique applied on other bit of LSB, also the PSNR of same embedding capacity that is applying on cover image is higher than embedding the same capacity on gray image, but the proposed technique is more concerned with the security of the handwritten signature. The results of the proposed technique is scaled using the same ratio of capacity to cover size in technique [17] to compare its results, since the authors of this technique used a colored cover image 3 levels (R,G,B) of size (512*512) and the proposed technique uses only one level (gray level) of size (512*512). Moreover, a proposed technique applied the 1st LSB on gray cover image to compare its results with technique [17], as this technique apply sequential LSB.

TABLE II. MSE OF THE PROPOSED AT DIFFERENT IMAGES DATASET (APPLYING ON HANDWRITTEN IMAGE) WITH GRAY COVER IMAGE AND EXISTING ALGORITHMS (APPLYING ON TEXT) WITH COLORED COVER IMAGE IN [17] AND A PROPOSED METHOD APPLYING 1st LSB ON GRAY COVER IMAGE

Cover Image	Cover Size	Capacity of message	Proposed MSE	MSE in [17]	proposed MSE of 1 st LSB
lena.jpg	512*512	2K	0.3395	0.0211	0.0327
pears.png	512*512	2K	0.3235	0.0214	0.0323
peppers.jpg	512*512	2K	0.2242	0.0212	0.0322

The results in these tables show that the PSNR values for the proposed LSB is high for embedding 2 K bytes in the gray cover image than using the technique in[17] that embeds capacity of 2000 bytes and embeds those bits in color covered image i.e (it embeds more than bits in one pixel unlike our proposed technique that embeds only one bit in the pixel, the PSNR values of that proposed LSB is high as Typical values for the PSNR and lossy media compression has normal value of 30 and 50 dB, where higher value is better [7] and the value of PSNR obtained by the proposed LSB also has the acceptable range. The main concern of this research is about security; the use of hybrid Cryptography RSA and proposed steganography LSB is more secure than using only steganography sequential LSB for using two parameters: randomizing in place of pixels for inserting the bit of handwritten signature and randomizing in selection of number of bit in that pixel to do bit scattering (not normal LSB, but random in range of last three bits).

TABLE III. PSNR OF THE PROPOSED METHOD AT DIFFERENT IMAGES DATASET (APPLYING ON HANDWRITTEN IMAGE) WITH GRAY COVER IMAGE AND EXISTING ALGORITHMS [17](APPLYING ON TEXT) WITH COLORED COVER IMAGE AND A PROPOSED METHOD APPLYING 1st LSB ON GRAY COVER IMAGE

Cover Image	Cover Size	Capacity of message	Proposed PSNR	PSNR in [17]	proposed PSNR of 1 st LSB
lena.jpg	512*512	2K	52.82	64.89	62.98
pears.png	512*512	2K	53.03	64.83	63.03
peppers.jpg	512*512	2K	54.62	64.88	63.05

B. Entropy:

A further parameter is also used to evaluate the cover image and stego image in table IV, i.e. Entropy (Average content for information). It tests the proportions of the picture's data. It is commonly calculated as bits in units.

$$Ent(p) = \sum_{i=0}^T Pro(i) \log Pro(i) \tag{9}$$

where $pro(i)$ is the function of a given image's probability density at intensity level l , and T is the total number of grey levels in the image. An picture with a high value of entropy is known as having better quality and high information. The entropy's values shown in table IV provide that the proposed technique stores more information at the cover images than the existing technique, as our model deals with a handwritten image as a secret data of large capacity due to using the RSA encryption before embedding it rather than using text of small capacity in [19].

TABLE IV. ENTROPY'S OF COVER IMAGE AND MODIFIED (STEGO) IMAGE

image of size (512*512)	Cover image's Entropy	modified image's Entropy	Cover image's Entropy in [19]	modified image's Entropy in [19]
lena.jpg	7.4482	7.4503	7.4469	7.4470
pears.png	7.2591	7.2600	7.2587	7.2588
peppers.jpg	7.5903	7.5918	7.5818	7.5819

TABLE V. PSNR OF PROPOSED METHOD AND OTHER TECHNIQUES BY HIDING 8KB OF DATA IN IMAGES OF (256*256).

Image name	Classic LSB method in [23]	Method in [20]	Proposed Method
Lena	42.51	49.37	45.22
Baboon	54.73	49.38	45.65
Pears	43.24	49.41	45.34

C. Comparison between Proposed method and other techniques

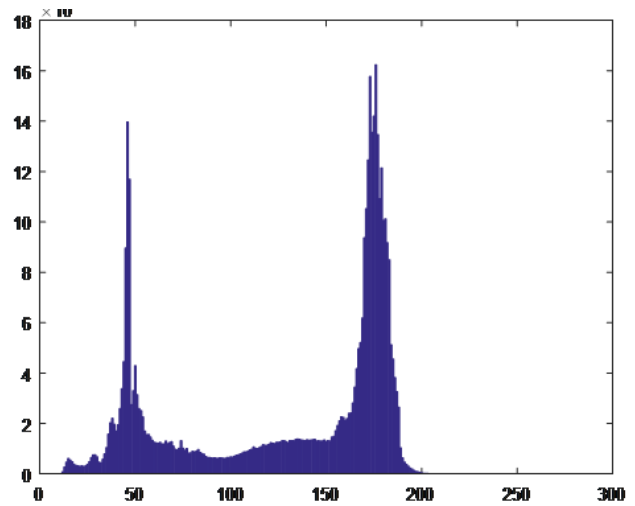
The value of PSNR for the proposed method is compared to that of various methods, with the results shown in Table V. The 8 KB message data (ciphered handwritten image in proposed method) is converted to binary and applied to standard images with a resolution of 256*256. Table V displays the effects of various techniques' PSNR values when applied to different images; the PSNR values for the other technique are taken from [23] and [20]. The LSB approach is simple to deconstruct. The method in [20] provides more capacity, but the proposed method is concerned with security more than capacity. The proposed method provides more security due to applying cryptography and hiding data in one of the last three bit randomly and not sequential.

D. Histogram analysis:

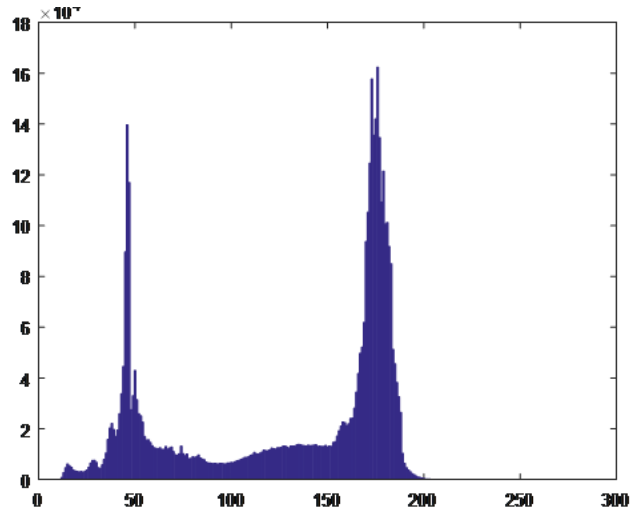
The image histogram is computed for the carrier media and the modified image and clarified in figures 7a, 7b, 8a, 8b and 9a, 9b. Where in figure 7a a carrier media is used having A5 paper size and it clarifies the histogram of carrier media at the left and the histogram of the modified image in figure 7b after applying the proposed technique of embedding at the right, in figure 8a a Pepper cover image of size 512*512 is used it shows histogram of covered image at the left and the histogram of the modified image in figure 8b after applying the proposed technique of embedding at the right and in figure 9a an Lena cover image of size 512*512 is used and it shows histogram of covered image at the left and the histogram of the modified image in figure 9b after applying the proposed technique of embedding at the right. The histograms of both the images (covered and stego) are quite similar where the histogram of proposed technique has no difference from the carrier media rather than the sequential LSB which there exist some difference between the carrier image and the modified image of that technique. Hence the proposed technique is found to be outperforming in comparison to existing techniques.

V. CONCLUSION

Signature is an important matter that is used in our daily life to accomplish any authentication for papers in work environment needing the signature of the person himself. There are many techniques for securing the data such as cryptography and steganography (like spatial domain, Transform domain). Cryptosystem and spatial domain is the area of interest in this research. The usage of the public key cryptosystem like (RSA) in hybridization with steganography provides more enhanced paradigm for securing the process of hiding human handwritten signature. In this research, we proposed such a paradigm that



(a) Carrier media



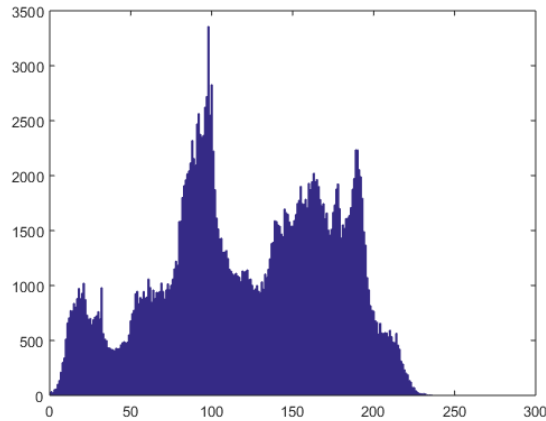
(b) Modified-Image

Fig. 7. Histogram Proposed Technique of LSB Embedding with Size A5.

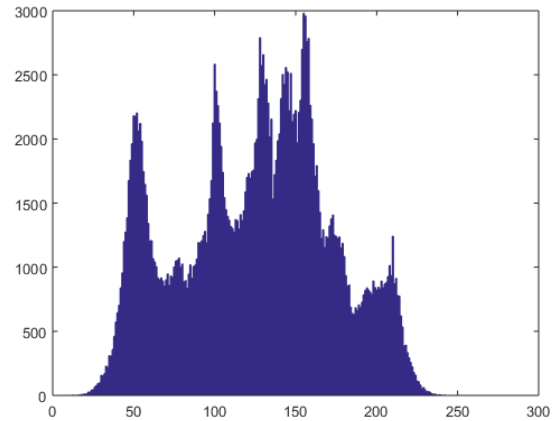
uses RSA in conjunction with scattering LSB based on pre-calculated random mathematical equation. The main concern of the proposed technique is about securing handwritten signature based on cryptosystem (RSA) and steganography, using modified LSB (scattered LSB in choosing pixel place and randomly choosing bit number to do LSB), rather than using sequential LSB. In contrast to the LSB method, our method does not consider its dependence on the 8th bit. One of the most important requirements of steganography is to embed the hidden message inside the carrier image without altering it significantly. Our method also meets this criterion to a higher degree. The experimental results show enhancement of PSNR value and histogram, the overall security of the scattered distribution provides an advanced security scheme comparing to the fixed allocation LSB.

REFERENCES

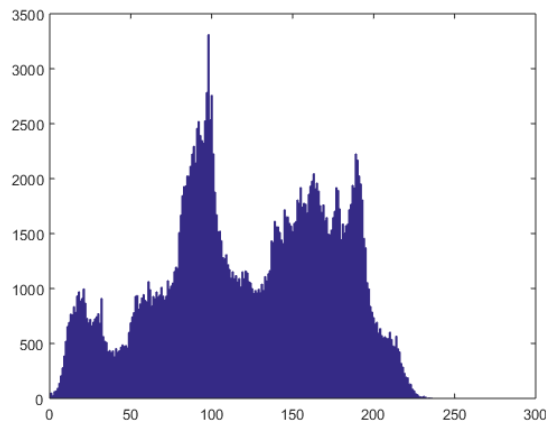
[1] D. P. M. Vijay Kumar Sharma, Dr. Devesh Kr Srivastava, "A study of steganography based data hiding techniques," *International Journal of*



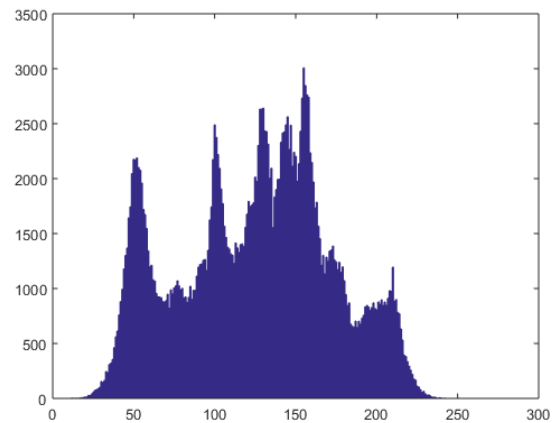
(a) Carrier media



(a) Carrier Media



(b) Modified-Image



(b) Modified-Image

Fig. 8. Histogram Proposed Technique of LSB Embedding with Pepper Cover Image of Size 512*512

Fig. 9. Histogram - Proposed Technique of LSB Embedding with Lena Cover Image of Size 512*512.

Emerging Research in Management and Technology, 2017.

[2] N. B. Dipti Kapoor Sarmah, "Proposed system for data hiding using cryptography and steganography," *International Journal of Computer Applications*, 2010.

[3] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, "A survey on the cryptographic encryption algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 11, pp. 333–344, 2017.

[4] M. S. Taha, M. S. M. Rahim, S. A. Lafta, M. M. Hashim, and H. M. Alzuabidi, "Combination of steganography and cryptography: A short survey," in *IOP conference series: materials science and engineering*, vol. 518, no. 5. IOP Publishing, 2019, p. 052003.

[5] S. M. H. Mohammad Ajman Hossain, "Steganography techniques: A review paper," *International Journal of Contemporary Computer Research (IJCCR)*, 2017.

[6] G. K. Rashmeet Kaur Chawla, "Comparative study on different steganographic techniques," *International Journal of Scientific Research and Management (IJSRM)*, 2015.

[7] F. A. O. Marwa M. Emam, Abdelmgeid A. Aly, "An improved image steganography method based on lsb technique with random pixel selection," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2016.

[8] F. A. O. Marwa E. Saleh, Abdelmgeid A. Aly, "Data security using

cryptography and steganography techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2016.

[9] T. S. S. S. Manisha, "A two-level secure data hiding algorithm for video steganography," *Multidimensional Systems and Signal Processing*, 2019.

[10] I. S. Brindha, "Hiding fingerprint in face using scattered lsb embedding steganographic technique for smart card based authentication system," *International Journal of Computer Applications*, 2011.

[11] A. K. G. Unik Lokhande, "Steganography using cryptography and pseudo random numbers," *International Journal of Computer Applications*, 2014.

[12] K. H. Shamim Ahmed Laskar, "Steganography based on random pixel selection for efficient data hiding," *International Journal Of Computer Engineering and Technology (IJCET)*, 2013.

[13] C. Kim, D. Shin, B.-G. Kim, and C.-N. Yang, "Secure medical images based on data hiding using a hybrid scheme with the hamming code, lsb, and opap," *Journal of Real-Time Image Processing*, vol. 14, no. 1, pp. 115–126, 2018.

[14] J. E. Anguraj S, Shantharajah S P, "A steganographic method based on optimized audio embedding technique for secure data communication in the internet of things," *Computational Intelligence*, 2019.

[15] E. A. K. Nouf A. Al-Juaid, Adnan A. Gutub, "Enhancing pc data security via combining rsa cryptography and video based steganography," *JOURNAL OF INFORMATION SECURITY AND CYBERCRIMES*

RESEARCH (JISCR), 2018.

- [16] N. A. Al-Otaibi and A. A. Gutub, "2-layer security system for hiding sensitive text data on personal computers," *Lecture Notes on Information Theory*, vol. 2, no. 2, pp. 151–157, 2014.
- [17] S. Chauhan, J. Kumar, A. Doegar *et al.*, "Multiple layer text security using variable block size cryptography and image steganography," in *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICCT)*. IEEE, 2017.
- [18] A. Singh and H. Singh, "An improved lsb based image steganography technique for rgb images," in *2015 IEEE International Conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2015, pp. 1–4.
- [19] S. Bhargava and M. Mukhija, "Hide image and text using lsb, dwt and rsa based on image steganography," *ICTACT Journal on Image & Video Processing*, vol. 9, no. 3, 2019.
- [20] K. Joshi, S. Gill, and R. Yadav, "A new method of image steganography using 7th bit of a pixel as indicator by introducing the successive temporary pixel in the gray scale image," *Journal of Computer Networks and Communications*, vol. 2018, 2018.
- [21] W. Stallings, *Cryptography and network security, 4/E*. Pearson Education India, 2006.
- [22] I. Vellore Institute of Technology University, "Handwritten signature dataset," 2018, accessed Feb 2019. [Online]. Available: <https://www.kaggle.com/divyanshrai/handwritten-signatures>
- [23] K. Muhammad, M. Sajjad, I. Mehmood, S. Rho, and S. W. Baik, "A novel magic lsb substitution method (m-lsb-sm) using multi-level encryption and achromatic component of an image," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14 867–14 893, 2016.

Design and Performance Measurement of Energy-based Acoustic Signal Detection with Autonomous Underwater Vehicles

Redouane Es-sadaoui¹, Jamal Khallaayoune²
Department of Electronic, Microwave and Optic
National Institute of Posts and Telecommunications
Rabat, Morocco

Tamara Brizard³
Arkeocean SARL, 808 route de Tourrette Levens
06790 Aspremont, France

Abstract—The Autonomous Underwater Vehicles (AUVs) industry is still awaiting its Henry Ford to bring to the market solutions that are well adapted to the challenge of underwater exploration. This will certainly be done by the advent of small connected drones equipped with small sensors and embedded devices, allowing AUVs to operate in a coordinated swarm, at a unit price so affordable that we can consider deploying hundreds, or even thousands simultaneously, to be able to observe the ocean with an instrument of a size finally adapted to its immensity. The scope of this work is to build a high performance and low-cost embedded device easy to mount onboard small AUVs and implementing energy-based spectrum sensing algorithms in order to detect targets underwater using acoustic waves. The principle of design, hardware architecture and real-time implementation of the acoustic signal processing chain are described in this paper. Simulations and sea experiments have been conducted successfully and qualified the performance of the realized system to detect acoustic pings underwater depending on the signal-to-noise ratio (SNR). Moreover, this paper proposes methods to improve the measured detection range and accuracy.

Keywords—Autonomous Underwater Vehicles (AUVs); acoustic signal processing; spectrum sensing; energy detection

I. INTRODUCTION

Many questions have been raised after the mission failure of BlueFin-21 (Fig. 1), a super-equipped AUV costing more than million dollars deployed in March 2014 by the U.S. Navy to spot black boxes of Boeing 777 of Malaysian Airlines flight MH370, crashed into the empty vastness of the southern Indian Ocean, killing all 239 passengers and crew onboard.



Fig. 1. From Left to Right : the Bluefin-21, Bluefin-12, and Bluefin-9 Autonomous Underwater Vehicles (AUVs) [1].

To successfully capture the last acoustic signals transmitted by the black boxes, it would probably have been necessary to deploy a multitude of coordinated "swarm", forming a listening network deployed in the column of water over several square kilometers (see Fig. 2). But with AUVs costing tens or even hundreds of thousands of dollars, this concept is economically unrealistic.

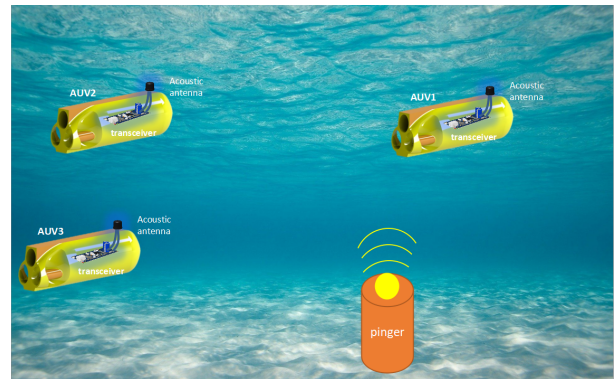


Fig. 2. Schematic Showing the Proposed Solution to Detect Targets Underwater using Small AUVs Operating in Swarm.

In recent years, as covered by [2], [3], oceanic based research is gradually growing across the globe trying to solve the major challenges of Underwater Acoustic Sensor Networks (UWASNs) like limited bandwidth, propagation, power constraint, localization and spectrum utilization, which is challenging compared to terrestrial applications due to the unavailability of GPS and to the unique physics of sensing in the marine environment where the search zone is immense, and the bad weather and rough seas hamper efforts to find objects underwater. However, the majority of research works are limited to theoretical studies and lack of real implementation with weak contribution in the technical development of AUV industry.

The motivation behind this work arises from an absence of open architectures and easy-to-use solutions that give access to the physical variables of underwater acoustic systems, which, in turn, will allow the research community to design novel acoustic spectrum sensing algorithms. In this scope, the present paper presents a proof of concept and major steps required to

design and develop a low volume acoustic device that can be installed onboard AUVs in order to detect targets underwater. These three main challenges will be investigated:

- 1) Addressing the "Underwater spectrum sensing challenge" : where the AUV should be able to sense acoustic signals transmitted by subsea targets.
- 2) Addressing the "Acoustic ranging challenge": providing the AUV with its relative distance to the transmitter of acoustic waves.
- 3) Addressing the "accurate detection challenge": Improving the detection performance especially in term of maximum range and Time-of-Arrival (ToA) measurement accuracy.

The paper is organized as follows. Section II reviews the acoustic spectrum sensing techniques. In Section III, the proposed methods are described by specifying the hardware architecture of the pinger and the transceiver, implementation of the acoustic signal processing chain and its simulation in MATLAB. Section IV presents the sea trial results and discusses solutions to improve the performance of the realized system. Conclusions are given at the end.

II. SPECTRUM SENSING METHODS REVIEW

A. Overview

Above water, most autonomous systems rely on radio or spread-spectrum communications and global positioning. However, electromagnetic signals attenuate rapidly underwater, and therefore acoustic waves are needed because they are less attenuated and travel further in water, as detailed in [4] and [5]. The use of electromagnetism remains possible at "very low" or "extra low" frequencies, but their implementation to establish a subsea communication will require large antennas in general, incompatible with the small size of an AUV and will at most allow a few bits per second to be exchanged, which remains too limited for the operational use that is generally required with AUVs.

The terms "Spectrum sensing" and "detection" are nominally interchangeable. They denotes the process to identify the presence or absence of transmitters in a specific spectrum; see[6]. Spectrum sensing techniques can be classified into two main categories :

- Non-cooperative [7], [8]: where receivers must independently have the ability to determine the presence or absence of a transmitter in a specific spectrum.
- Cooperative [9], [10], [11]: where a group or network of receivers exchange information in order to enhance the detection performance.

This paper focuses on Non-cooperative category allowing each AUV to detect and recognize transmitters pinging in a specific spectrum. For reason of simplicity, the following comparison focuses on the two popular techniques (covered here [12]) that are matched-filtering and energy based detection.

B. Matched Filtering based Detection

The matched filtering based spectrum sensing technique, as detailed in [13], [14], uses the transmitted signal as a

template to which the received signal is compared. The better match between the template and the received signal, is the greater amplitude of the matched filter's output. The received signal and pilot signal are convoluted and averaged over N samples to obtain the matched filter decision statistic, which is then compared to the matched filter threshold T_{MF} to get the sensing decision. T_{MF} is calculated by:

$$T_{MF} = \frac{1}{N} \sum_{n=1}^N (y(n) * x_p(n)) \quad (1)$$

Where the received signal stream is denoted by $y(n)$, the known pilot signal is indicated by $x_p(n)$, and N is the number of samples acquired in a sensing cycle Under H_0 , the decision statistic, T_{MF} , is obtained by the averaged convolution of the Gaussian noise and the pilot signal. On the other hand, under H_1 , T_{MF} results from the convolution of the transmitted signal contaminated with the Gaussian noise and the pilot signal averaged over N samples. The matched filter threshold, λ_{MF} is taken from the "quiet time approach". Therefore, the noise is merely present in the received signal, $y(n)$. As a result, the matched filter threshold, λ_{MF} is identical to the matched filter decision statistic, T_{MF} during the quiet time period. If λ_{MF} is determined, the binary hypothesis is given as:

$$\begin{aligned} H_0 : T_{MF} &> \lambda_{MF} \\ H_1 : T_{MF} &< \lambda_{MF} \end{aligned} \quad (2)$$

C. Energy based Detection

Energy detection (also called non-coherent detection; in reference to [15], [16]) is very popular and performed by simply comparing the output of the energy of the received signal energy with a predefined threshold . The decision statistic of an energy detector can be calculated from the squared magnitude of the FFT averaged over N samples of the received signal. The detector output is the received signal energy as given by:

$$T_{ED} = \sum_{n=0}^N y(n)^2 \quad (3)$$

Where $n = 1 \dots N$, N is the number of samples, and $y(n)$ is the received signal, and T_{ED} denotes the energy of the received signal. The detection decision can be expressed as:

$$\begin{aligned} H_0 : T_{ED} &> \lambda_{ED} \\ H_1 : T_{ED} &< \lambda_{ED} \end{aligned} \quad (4)$$

Where λ_{ED} denotes the energy detection threshold.

Our design will be based on energy detection mechanism which is easy to implement with moderate computational complexities and can be performed on both time and frequency domain. In addition, compared to the matched filtering, Energy detection does not require a prior information of the transmitted signal to operate. However, the detection threshold has to be selected carefully, as described in the next section.

D. Detection Threshold

The detection threshold is a decibel number that essentially incorporates the AUV acoustic transceiver ability to decide

that a detection is made or not made. The detection process generally includes the following probabilities:

- The probability of detection (Pd): the probability that a signal is detected if it is present;
- 1-PD: the probability the signal will not be detected if it is present;
- The probability of false alarm (PFA): the probability that a signal is detected when it is not present;
- 1-PFA: the probability that the signal will not be detected when it is not present

As described in [17], [18], the detection threshold DT with a Gaussian signal for a given Pfa :

$$DT = 10 \log \frac{Q^{-1}(Pfa) - Q^{-1}(Pd)^2}{2} \quad (5)$$

Where $Q(\lambda_{ED}) = Proba(Z > \lambda_{ED})$ is the error function that validates the probability where a gaussian signal Z of variance σ and average M reaches a threshold λ_{ED} .

The problem exposed here corresponds to a frequent question in a detection chain: how to adjust the threshold to control the false alarm rate? In the ideal case, the statistics of the noise alone are known and the threshold can be calculated (analytically or by simulation) from the desired false alarm rate, but in practice the characteristics of the noise are variable and it is therefore necessary to estimate the noise in permanently to have an correct estimation of the threshold. We consider here the case of a quadratic detection chain fed by Gaussian noise, the variance of which is known as detailed in [17], [18]. In this case the decision variable consists of the square of the input Gaussian variable and is therefore distributed according to an exponential law of density :

$$p(x) = \frac{1}{2\sigma^2} \cdot \exp \frac{-x}{2\sigma^2} \cdot U(x) \quad (6)$$

Where $U(x)$ designates the function which is worth 1 for x positive and 0 elsewhere (HEAVISIDE [19]). We can easily set the detection threshold from the desired Pfa , for example, which $Pfa = 10^{-10}$ is not abnormal in an automatic detection device. We have :

$$Proba(X > \eta) = \exp \frac{-\eta}{2\sigma^2} = 10^{-10} \quad (7)$$

Thus,

$$\eta = 10 \cdot \ln 10 \cdot 2\sigma^2 \quad (8)$$

We can therefore write the detection condition in the form:

$$X - 10 \cdot \ln 10 \cdot 2\sigma^2 > 0 \quad (9)$$

In practice $2\sigma^2$ is unknown and must therefore be replaced by an estimator noted Y . The detection condition then becomes:

$$X - \lambda Y > 0 \quad (10)$$

As result, the detection threshold will be biased and the value of λ will be different from $10 \cdot \ln 10$ due to the presence of Y instead of $2\sigma^2$.

III. PROPOSED METHODS

A. Acoustic Pinger Design

Refer to Fig. 3 for the next discussion. The acoustic pinger incorporates a PIC18 Micro-controller (MCU) [20] offering high computational performance at an economical price – with the addition of high-endurance, Flash program memory and introducing design enhancements that make these Microcontrollers a logical choice for many high performance, power sensitive applications. The Flash MCU outputs acoustic pulses in 8-bits digital format, the pulse is converted to an analog format by a Digital to Analog Converter (DAC), then amplified to high amplitude (around 24 volts) by means of an analog converter. A transformer is installed after the amplifier to raise the voltage of the acoustic pulse to a high amplitude voltage, which will attack an acoustic transducer ceramic that convert the electrical wave to an acoustic wave propagating in the water.

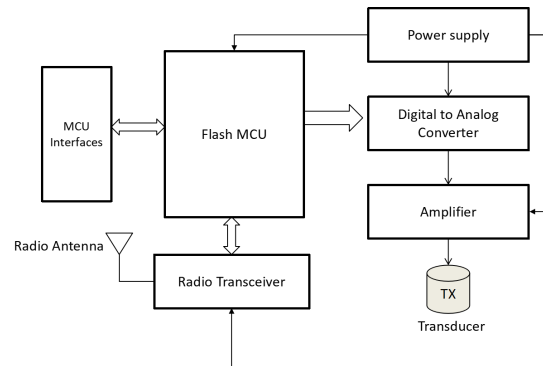


Fig. 3. Schematic Diagram Showing the Acoustic Pinger Architecture.

The DAC circuit was designed to interface between PIC MCU that are generating the acoustic pulse in digital format and the analog amplifier. It is rated as single, 8-bit, voltage out DAC that operates from a single 2.7 to 5.5 Volts supply and has a parallel microprocessor and DSP compatible interface with high speed registers and double buffered interface logic. Its on-chip precision output buffer allows the DAC output to swing rail to rail. The low power consumption of this part makes it ideally suited to portable battery operated equipment. The transmitter amplifier power supply is boosted to 24 Volts.

B. Acoustic Transceiver Design

The transceiver boards presented in Fig. 4 contains all the necessary circuits to condition and process signals received by the acoustic hydrophone.

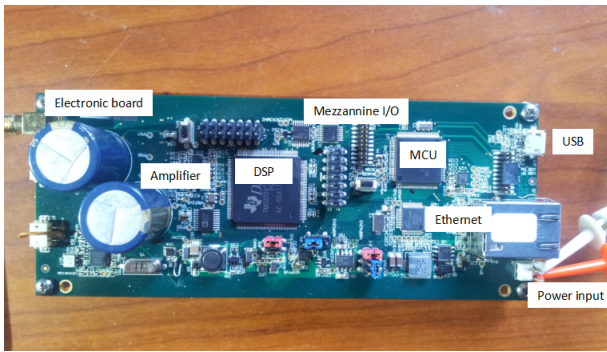


Fig. 4. Picture of the Acoustic Transceiver Board.

This section scrutinizes the blocks from Fig. 5, as well as their relations. The transceiver board can receive its power supply from the AUV and operate from 3.7 to 14 Volts. The power consumption of the board is about than 60 mA when fed under 12 Volts. The option to have an additional separate 24 Volts supply directly feeding the acoustic transmitter is also offered. With this option, the two large electrolytic capacitors are no longer necessary. A mezzanine for custom IOs or functions can be installed on the board as a MCU peripheral. many wireless on-chip modules were integrated through the mezzanine allowing the board to communicates via Radio 868MHz / 912MHz, Lora, or WiFi.

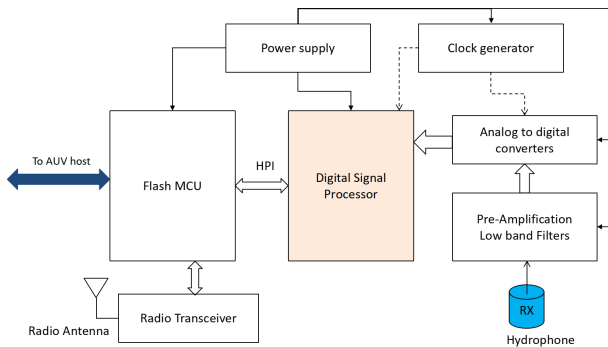


Fig. 5. Schematic Diagram Showing the Acoustic Transceiver Architecture.

The board incorporates a Texas Instruments TM320C5x Digital Signal Processor (DSP) [21] and MSP430F5xx MCU [22]. The DSP is a fixed-point processor based on an advanced modified Harvard architecture which provides an arithmetic logic unit (ALU) with a high degree of parallelism, application-specific hardware logic, on-chip memory, and additional on-chip peripherals. The MSP430F5 family features a powerful 16-bit RISC CPU, 16-bit registers, and constant generators that contribute to maximum code efficiency. These MCUs include a high-performance 12-bit analog-to-digital converter (ADC), up to four universal serial communication interfaces, hardware multiplier, DMA, real time clock module with alarm capabilities, and I/O pins. Onboard the transceiver, this MCU manages the RF transceiver, the communication to the DSP using Host Port Interface (HPI), and communication to an external host and sensors. Furthermore, this MCU supports TI RTOS based software and open to custom applications. Inter-processor communications were implemented between

the Flash MCU and the DSP based on the Host Port Interface link (HPI): A high speed parallel port through which the Flash MCU can directly access to DSP memory space.

Before being digitally processed, the hydrophone acoustic signals are first passed through a low pass filter and converted to digital format using a stereo ADC, that perform sampling, analog to digital conversion, and anti-alias filtering.

C. Acoustic Chain

As illustrated in Fig. 6, each received signal (a pure tone sine wave of constant frequency) is first amplified and filtered, then sampled through the ADC converter. The time-of-arrival of incoming pulses on the transceiver antenna is mainly measured by the DSP which implements an energy detector testing continuously if the pulse is present or not, and a high precision timer marking the detection timestamp. If we consider that the pinger and the transceiver share the same time reference [23], [24], the ToA of the incoming pulses can be estimated and the distance between the pinger and the transceiver will be determined.

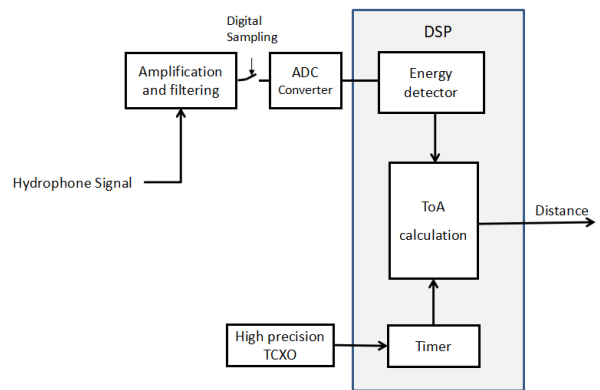


Fig. 6. Block Diagram Showing the Implementation of the Acoustic Signal Processing Chain. The Output is the Distance Between the Pinger and the Transceiver.

The transmitted pulse is a pure tone sine wave pulse, shaped in a balckman window: The pulse is combined from a blackman window multiplied by a sine wave. Equations (11) and (12) define respectively the mathematical formulas of the Blackman window and the transmitted pulse [25].

$$w(k) = 0.42 - 0.5 \cos\left(\frac{2\pi \cdot k}{N-1}\right) + 0.08 \cdot \cos\left(\frac{4\pi \cdot k}{N-1}\right) \quad (11)$$

Where $0 \leq k \leq M-1$, N is the length of the blackman window. M is $N/2$ when N is even and $(N+1)/2$ when N is odd.

$$S(k) = A \cdot \sin(2\pi \cdot f / f_e) \cdot w(k) \quad (12)$$

Where A is the signal amplitude peak, f is the signal frequency, f_e is the sampling rate. The acoustic pulse form was simulated in Matlab and implemented in the real platform. The real output of the transmitter was measured and qualified by using an oscilloscope. Fig. 7 shows the used waveform, for a signal frequency of 22 KHz, a sampling frequency of 48 KHz and a pulse length of 12 milliseconds.

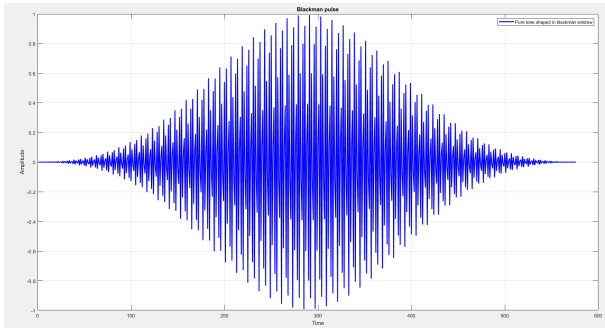


Fig. 7. Hydrophones Input Signals. Pure Tone Sine Wave Shaped in 12 ms blackman Window Without Noise. SNR = 100 dB, Frequency = 22 kHz, Amplitude = 20 mV, Pulse width = 12 ms, Sampling Rate = 48 kHz. Horizontal Axis is the Time in Samples (20.8 μ s). Vertical Axis is the Amplitude in Volts.

This pulse was implemented on the Flash MCU of the acoustic pinger. Oscilloscope measurement of the pulse at the output of amplifier is illustrated in Fig. 8.

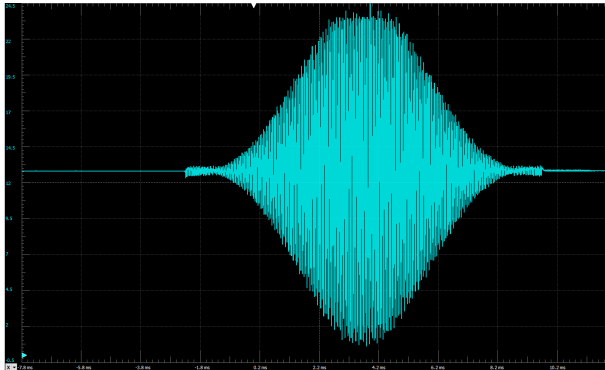


Fig. 8. Pulse Measured at the Output of the Acoustic Pinger Amplifier. Oscilloscope Configuration: Vertical Axis is the Amplitude in Volts (1.5V/div). horizontal Axis is the Time (2ms/div).

D. Energy Detector

The energies of the hydrophone channel can be calculated using the following formula:

$$Energy = \sum_{n=0}^N y_h(n)^2 \quad (13)$$

Where y_h is the received signal from the hydrophone and N is the number of samples.

1) *Pulse energy*: The short integrator is set to measure the pulse energy during a short period. By analogy with an RC filter [26], the equation of an RC integrator is as follows:

$$Y(n) = \frac{X(n)}{\alpha} + Y(n-1), RC_{\alpha} = \alpha.Te \quad (14)$$

The z transfer function of this integrator is:

$$H(z) = \frac{1}{\alpha - (\alpha - 1).Z^{-1}} \quad (15)$$

This is a first order low pass filter and cutoff frequency:

$$F_c = \frac{1}{2.\pi.RC_{\alpha}} \quad (16)$$

The short Integrator is updated with a new energy sample as follows:

$$IntgC(n) = \frac{Energy_{antenna} - IntgC(n-1)}{RC_{\alpha}} + IntgC(n-1) \quad (17)$$

2) *Noise energy*: The long integrator aims at estimating the noise level in a long period. Similar to the short integrator, the long integrator is updated with a new energy sample as follows:

$$IntgL(n) = \frac{En - IntgL(n-1)}{RC_{\beta}} + IntgL(n-1) \quad (18)$$

3) *Detection contrast*: The detection contrast can be then estimated:

$$Detection_Contrast = \frac{IntgC(n)}{IntgL(n)} \quad (19)$$

E. MATLAB Simulation

The acoustic chain and energy detector described above were implemented in MATLAB. The objective of this simulation is to qualify the capability of the designed chain to detect acoustic properly. The test signals are considered as pure tone sine waves shaped in blackman window and combined with white gaussian noise. Fig. 9 shows the frame used for the test, which is combined from pulses with 200 ms time-spaced simulating the reception of acoustic pulses by the transceiver.

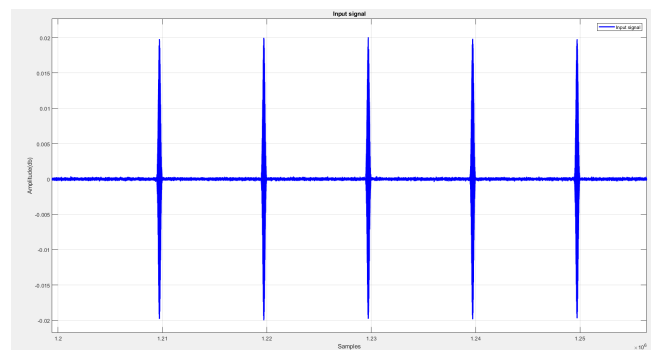


Fig. 9. Acoustic Pulses Frame Showing the form of Pulses Received by the Transceiver. SNR = 40 dB, Frequency = 22 kHz, Amplitude = 20 mV, Pulse width = 12 ms, Sampling Rate = 48 kHz. Horizontal Axis is the Time in Samples (20.8 μ s). Vertical Axis is the Amplitude in Volts.

Fig. 10 gives the received pulses pass through a low pass filter of 330 coefficients (6.8 ms). This filter is able to attenuate adjacent channels with up to -80 dB at 500 Hz which allows the acoustic chain to process multi-channels pulses in the range between 18 and 22 kHz.

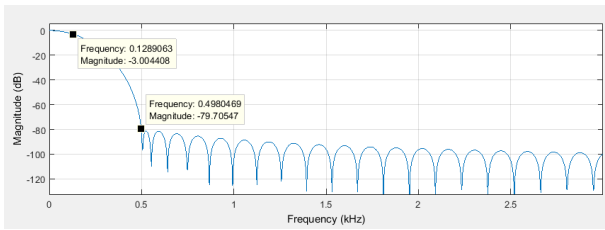


Fig. 10. Low Pass Filter Design in MATLAB. 330 Coefficients (fixed-point conversion Q21), Cutoff Frequency at -3db: 129 Hz, Adjacent Channel Attenuation : -80 dB at 500 Hz.

Fig. 11 gives a comparison of the integrators measurements in response to the acoustic frame. In blue is the 5.33 ms short integrator following the shape of the pulse. The long integrator, in red, gives the estimation of noise.

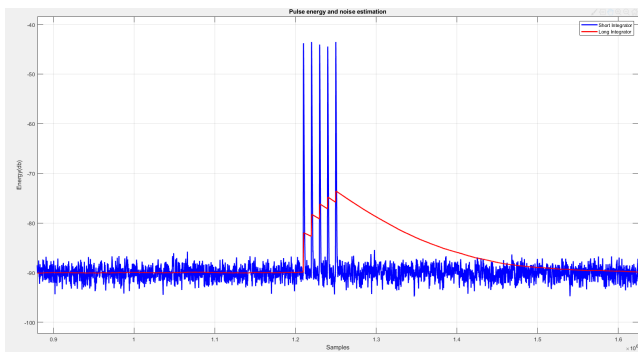


Fig. 11. Integrators Measurements. Horizontal Axis is the Time in Samples (20.8 μ s). Vertical Axis is the Amplitude in dB.

The energy detection profile for the existing system (IntegL = 1366 ms, IntegC = 5.33 ms, 12 dB detection threshold) is also presented in Fig. 12. The blue solid line is the 12 dB detection threshold. The energy detection level, in red color, ($= \frac{IntgC}{IntgL}$) exceeds the detection threshold when pulses are received.

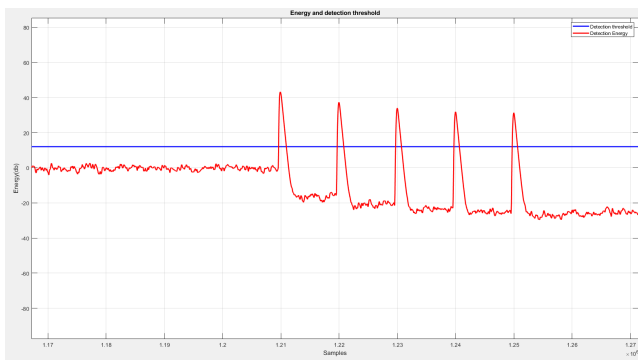


Fig. 12. Energy Detection Profile of the Acoustic Chain. Horizontal Axis is the Time in Samples (20.8 μ s). Vertical Axis is the Amplitude in dB.

IV. RESULTS AND DISCUSSION

A. Sea Experiments

This section presents two sets of experimental results. The developed system was first tested in the Marina Bouregreg harbor, located at the mouth of the Bouregreg River, on the shore of SALE, Morocco. Then we reproduced tests in Guerlédan lake, France (Fig. 13).

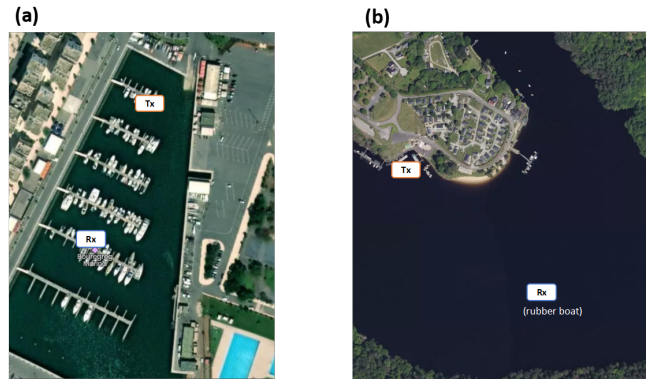


Fig. 13. Map Images Showing the Location of the Pinger (Tx) and the Transceiver (Rx) Installed in (a) Marina Harbor and (b) Guerlédan Lake.

1) *Experimental platform:* A schematic diagram of the experimental platform is shown in Fig. 14. The experiments reported herein rely on a transmitter electronic board that was set on the deck, while its watertight ceramic was immersed at 2 meters depth. The receiver electronic board, connected to a Laptop, was also set on the deck, with a distance to up to 280 meters far away from the transmitter. The receiver hydrophone antenna was immersed at a depth 2 meters. The transmitter and the receiver were put in line of sight.

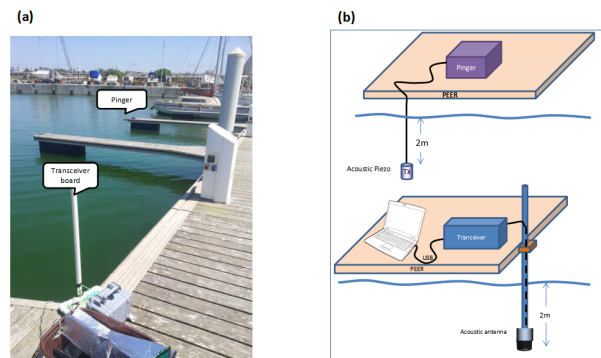


Fig. 14. (a) Image of the Experimental Platform at Marina Harbor. (b) Schematic Diagram Showing the Acoustic Pinger and the Acoustic Transceiver Boards.

2) *Results:* The acoustic pinger was programmed to send 20 kHz pure tone acoustic pulses at every second. The acoustic transceiver is receiving acoustic pulses through the acoustic antenna immersed in water. The digital signal processor detects pulses (at 20 kHz frequency) and computes the time of arrival. The Flash MCU of the transceiver reads data from DSP

memory through HPI link and outputs the computed distance to a laptop through serial USB. The sea trial results have been obtained by first synchronizing the pinger and the transceiver by radio to have the same time reference. Then by varying the distance between the pinger and the transceiver and measuring the time-of-arrival (ToA) of acoustic pulses at the transceiver. The recorded ToAs are logged into the Laptop PC in reference to the appropriate position. The DSP firmware parameters were set to: (a) Short Integrator period of 5.33 ms. (b) A Long Integrator period of 1.366 s. (c) Detection threshold of 12 dB. (d) The sound velocity was taken equal to 1500 m/s.

Fig. 15 shows an example of experiments data recorded in Guerlédan lake with pinger deployed 45m far away from the transceiver. This graph shows a minimum error of 0.003m, a maximum error of 1.48m, an average error of 0.62m and a peak-to-peak variance of 1.47m.

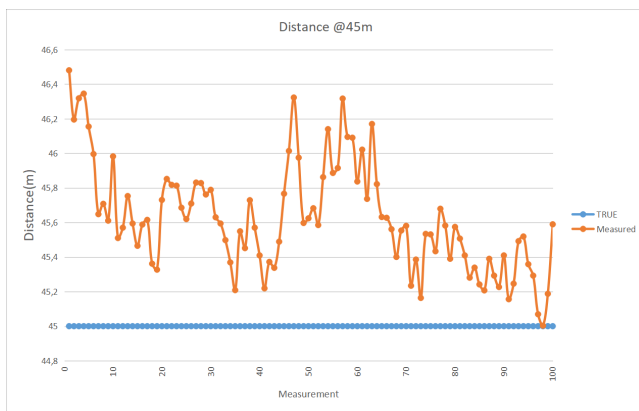


Fig. 15. Measured Distance at 45m. In Blue is the True Value (45m). In Orange is the Measured Distance by the Transceiver. Horizontal Axis is the Measurement Number. Vertical Axis is the Distance in Meters.

Table I summarizes the obtained performances. The detection contrast at Marina harbor conditions (low depth, noisy environment) was estimated to be very low (less than 20 dB at 280m), which explains the low performance (520m maximum range, +/-9m error at 280m). A better performance was obtained (1.2 km detection range, +/-4m error at 300m) when we reproduced the same experiments in the Guerlédan lake (France) with better conditions: the detection contrast was measured around 32 dB at 300 meters.

TABLE I. SEA EXPERIMENTS PERFORMANCE. PINGER PULSE WIDTH = 12 MS, SAMPLING RATE = 48 KHZ. TRANSCEIVER DETECTION THRESHOLD = 12 DB

Environment	Water depth	Distance accuracy	detection range
Marina harbor	4m	+/- 9m measured at 280m	520m
Guerlédan lake	30m	+/- 4m measured at 300m	1.1 km

B. Improving the Noise Estimation

1) *Observation:* During sea trial experiments, we have noticed that the channel noise measurements given by transceiver onboard the vehicle considerably increase with the reception of acoustic pulses from transmitters as presented in Fig. 16.

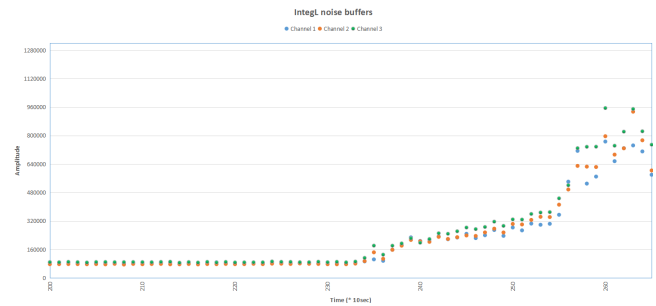


Fig. 16. Long Integrator Noise Buffer Recorded During Sea Trial Experiments at Guerlédan Lake, France. Vehicle Navigating in Free Run. Acoustic Pulses Received Every 200 ms.

This obtained records show that the presence of the signal in the acoustic chain leads to an overestimation of the noise which has a significant “blanking” effect. Furthermore, the chain has a memory that is intentionally adjusted to obtain a correct estimate of the noise (a long memory). This phenomena can be proven using formulas in Section ?? . The present context is rather favorable since the expected signal is known. We can therefore inhibit the calculation of the variance of noise during the pulse presence: (1) Either by delaying the long integration calculation quite enough to make sure the signal does not pollute the noise. (2) Or by simply freezing the long integrator at the first sign of the presence of the signal, as described in the next section.

2) *Proposed solution:* Refer to Fig.17 for the next discussion. The algorithm of freezing the long integrator (IntgL) is proceeded as follows: In the absence of pulses, the Long Integrator is calculated as before where the short integrator (*IntgC* = 5.33ms) is considered to check the evolution of pulses energies. If the difference between the IntgC and the IntgL exceeds 3 dB, we stop feeding the IntgL. These simulations were made with an IntgL always frozen if the difference between the IntgCp and the IntgL is more than 3 dB, and de-frozen if the difference is less than 3 dB. Fig. 18 shows the new noise profile (Long integrator).

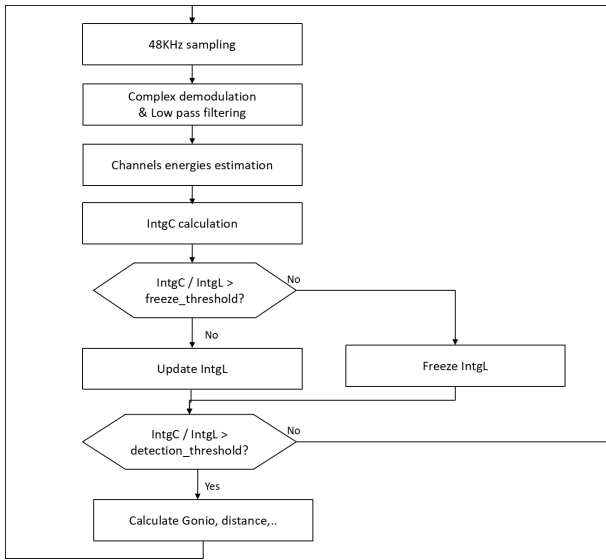


Fig. 17. Freezing IntgL Flowchart. Proposed to Remove the Sea Noise Estimation bias Caused by the Presence of the Pulses.

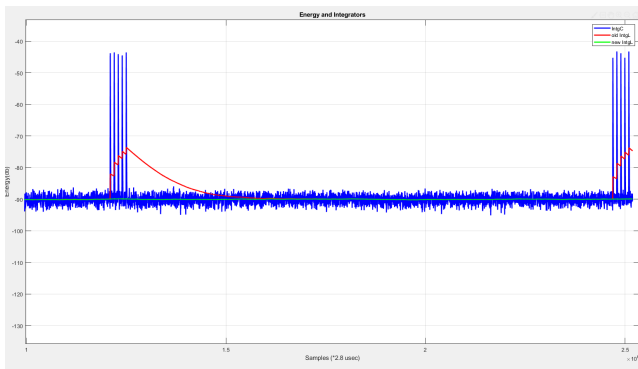


Fig. 18. Long Integrator Profile. In Blue, is the 5.33 ms Short Integrator. In Red, the Old Long Integrator with Value Increased in the Middle of Pulses. The New Long Integrator (in green) Maintains the Noise Level and Got Frozen when Pulses are Presents.

From the obtained results, we notice that the long integrator is sensible to the acoustic pulses reception which makes the estimation of noise level biased with an offset of more than 10 dB after some seconds of navigation. Means that the real detection threshold is biased with 10 dB and becomes 22 dB instead of 12 dB. The existing design where the long integrator is fed in permanence is working well with systems receiving pulses with low frequency (1 Hz or less). Otherwise, if we receive continuous pulses every 200 ms (5Hz or more), this conducts to a biased estimation of the noise level because of the high presence of pulses in noise estimation. So the way we are handling the long integrator is not optimal anymore. This was fixed by freezing the long integrator during the presence of the pulse which will increase the SNR and improve the detection range.

C. Improving the Distance Accuracy

As reported in the sea trial results shown in I, a limitation of the realized system appears in term of inaccuracy of ranging

(up to +/-9m were measured with low SNR configuration). This is explained by the fact that the existing chain is not optimized for accurate distance measurement because of the large width of the acoustic signal (12 ms). The objective of this section is to review the existing chain and propose solutions to improve its distance accuracy. Basically, decimetric accuracy on the distances is obtained by reducing the width of the acoustic pulses [27]: (1) Either by physical means, by shortening the transmission time in water or (2) by pulse compression methods (not studied in this paper). The following solutions were proposed for investigation: (1) The existing chain : First, we need to qualify the distance accuracy of the the classic processing chain with 12 ms blackman pulses that turns today in the transceiver. (2) An experimental processing chain: based on the existing chain, with implements a tiny square pulse of 208 μ s width instead of blackman pulse. In addition to increasing the sampling rate to to 96 kHz, we will need to adapt the short integrator to this pulse width. At this stage, we will keep the low pass filtering as is. (3) A high resolution processing chain: identical to the experimental chain, but with a low pass filtering well adapted to the width of the pulse.

1) *Qualification of the existing chain:* MATLAB simulations have been performed using a simple frame of pure tone acoustic pulses with 2 s time-spaced. Each pulse is a blackman sine wave combined with white gaussian noise as shown previously in Fig. 7. The measurement of detection ToA was performed by varying the energy of the input signal from high SNR (more than 100 dB) to low SNR (12 dB). An example showing the ToA of an acoustic pulse of 67 dB is illustrated in Fig. 19. The status of detection goes high when the contrast (difference between the pulse energy and the noise) reaches threshold of 12 dB.

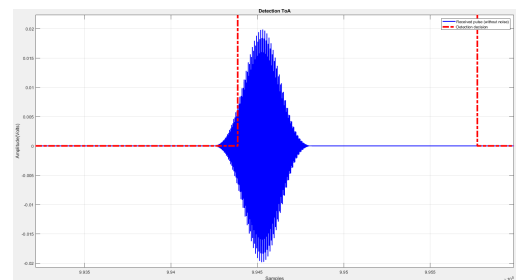


Fig. 19. Detection ToA Measurement with the Existing Chain. In Blue is the Acoustic Pulse. In Red is the Status of Detection (0=no detection, 1= detection). Blackman Pulse width = 12 ms, Sampling Rate = 48 Khz, Short Integrator = 6 ms, Low Pass Filter = 330 Coefficients. Horizontal Axis is the Time in Samples (20.8 μ s). Vertical Axis is the Amplitude in Volts.

The detection ToA and delays measurements, summarized in Table II, were obtained depending the variation of the detection contrast (SNR). The measurements show that the detection depends on the contrast level : with high SNR, the detection is at the beginning of the pulse, while the detection time is delayed to up to 8.916 ms (13,375 m distance accuracy) at the limit of detection (12 dB).

TABLE II. EXISTING CHAIN DETECTION DELAY MEASURED AS FUNCTION OF THE CONTRAST. BLACKMAN PULSE WIDTH = 12 MS, SAMPLING RATE = 48 KHZ, SHORT INTEGRATOR = 6 MS, LOW PASS FILTER = 330 COEFFICIENTS

Contrast (dB)	Detection delay (ms)	Distance accuracy (m)
more than 100 dB	reference	-
87 dB	1,458 ms	2,187 m
67 dB	2,354 ms	3,531 m
47 dB	3,583 ms	5,374 m
27 dB	5,625 ms	8,437 m
17 dB	7,354 ms	11,031 m
12 dB	8,916 ms	13,375 m

2) *Qualification of the experimental processing chain:*
 Similar to the qualification of the existing chain, the pulse frame contains acoustic pulses with 2 seconds time-spaced. However, the format of the pulse was to square sine wave instead of blackman, and the pulse width was reduced to 208 μ s instead of 12 ms. Fig. 20 illustrates the pulse format used and a ToA measurement when the SNR is around 82 dB.

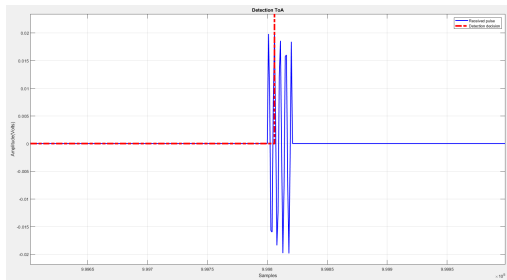


Fig. 20. Detection ToA with the Experimental Processing Chain. In Blue is the Acoustic Pulse. In Red is the Status of Detection (0=no detection, 1= detection). Square Sine Pulse width = 208 μ s, Sampling Rate = 96 KHz, Short Integrator = 96 μ s, Low Pass Filter = 330 Coefficients (3.4 ms). Horizontal Axis is the time in Samples (10.4 μ s). Vertical Axis is the Amplitude in Volts.

The detection ToA and delays of the experimental processing chain were measured depending the variation of the detection contrast level. The obtained results are presented in Table III. The measurements show that the maximum detection delay was reduced to up to 1.145 ms (around 1,718 cm distance accuracy compared to 13,375 m obtained previously with the existing processing chain).

TABLE III. EXPERIMENTAL PROCESSING CHAIN DETECTION DELAY MEASURED AS FUNCTION OF THE DETECTION CONTRAST. SQUARE SINE PULSE WIDTH = 208 μ s, SAMPLING RATE = 96 KHZ, SHORT INTEGRATOR = 96 μ s, LOW PASS FILTER = 330 COEFFICIENTS (3.4 MS). HORIZONTAL AXIS IS THE TIME IN SAMPLES (10.4 μ S). VERTICAL AXIS IS THE AMPLITUDE IN VOLTS

Contrast (dB)	Detection delay (ms)	Distance accuracy (m)
more than 100 dB	reference	-
82 dB	0,052 ms	0,078 m
62 dB	0,187 ms	0,281 m
42 dB	0,437 ms	0,656 m
22 dB	0,822 ms	1,234 m
12 dB	1,145 ms	1,718 m

3) *Qualification of the high resolution processing chain:*
 Now, with the high resolution processing chain, we are going

to adapt the filter window to the tiny pulse. By analogy to the low pass filter of the existing chain, the filter window was taken as the half of pulse width ($\frac{208}{2} = 104 \mu$ s). Which is set with 10 coefficients at 96 KHz sampling rate. Fig. 21 illustrates the low pass filter profile in MATLAB Filter Design Tool [28]. This filter is of course not very selective and not robust against adjacent channels, nevertheless it allows to reject out-of-band noises. Therefore, the high resolution processing chain will be limited to one-channel instead of eight.

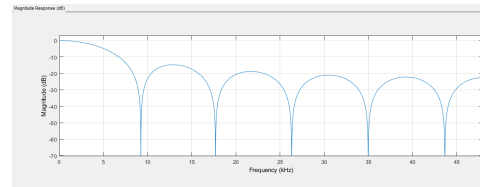


Fig. 21. Low Pass Filter Simulated with the High Resolution Processing Chain. 10 Coefficients FIR. Sampling Rate = 96 KHz. Rejection -15 dB at 7 KHz.

The detection ToA and delays of the high resolution processing chain were measured depending the variation of the contrast level. The obtained results are presented in Table IV. A very good performance was obtained with a maximum detection delay of 72.91 μ s (around 10,93 cm distance accuracy) at the limit of detection (12 dB).

TABLE IV. ONE-CHANNEL HIGH RESOLUTION PROCESSING CHAIN DETECTION DELAY MEASURED AS FUNCTION OF THE DETECTION CONTRAST. SQUARE SINE PULSE WIDTH = 208 μ s, SAMPLING RATE = 96 KHZ, SHORT INTEGRATOR = 96 μ s, LOW PASS FILTER = 10 COEFFICIENTS (104 μ S)

Contrast (dB)	Detection delay (ms)	Distance accuracy (m)
more than 100 dB	reference	-
86 dB	0 ms	0 m
66 dB	0 ms	0 m
46 dB	0 ms	0 m
26 dB	0,052083 ms	0,03124 m
12 dB	0,072916 ms	0,1093 m

V. CONCLUSIONS

In this paper, we presented the design of device able to detect acoustic signals underwater. We built prototypes that can be mounted onboard AUVs allowing them to detect and find targets that are sending acoustic waves. The proposed hardware architecture of the pinger is based on an ultra-low-power Flash Micro-controller sending pure tone acoustic pulses at narrow band frequency. The transceiver hardware onboard the AUVs incorporates a Digital Signal Processor implementing energy based spectrum sensing mechanism to detect the acoustic pulses sent by the pinger.

The acoustic chain was designed to process pure tone sine waves shaped in 12 ms blackman window at 48 kHz sampling rate, 6 ms short integrator and 330 coefficients (6.8 ms) of low pass filter. In addition, it supports multi-channels operation with the capability to recognize pulses from eight different pingers transmitting at different frequencies. The time-of-arrival of incoming pulses is then measured and the pinger position can be estimated.

Experiments with the realized system were carried out at sea with two different configurations: (a) at Marina harbor (considered as noisy environment), a very shallow waters (maximum depth ≤ 5 m) where the transceiver was able to detect pingers at 520m range, (b) at Guerlédan lake (considered as clean environment) with up to 30 m depth where more than 1 km range was obtained. During sea trials, two limitations of were identified: (1) The noise estimator was biased with more than 10 dB in the case where the receiver detects continuous pulses with a rate of 200 ms or less. We fixed this issue by freezing the long integrator at the beginning of pulse detection, which removed the bias and improved the detection range. (2) The distance accuracy was evaluated around 2 m in high SNR and 14 m at limit of detection. This performance can be improved to around 10 cm by using the one-channel high resolution processing chain with tiny square pulses of 208 μ s processed at 96 kHz sampling rate. As a consequence, the system will lose the feature of multi-channels operation. For future works, pulse compression technique should be investigated with matched filtering in order to enhance the performance of our system, especially in noisy environments.

ACKNOWLEDGMENT

This work was supported by Arkeocean SARL, France. The authors would like to thank the Marina harbor office who supported the sea trials of this work.

REFERENCES

- [1] Yao Yao. Cooperative navigation system for multiple unmanned underwater vehicles. *IFAC Proceedings Volumes*, 46(20):719–723, 2013. 3rd IFAC Conference on Intelligent Control and Automation Science ICONS 2013.
- [2] Ian F. Akyildiz, Dario Pompili, and Tommaso Melodia. Underwater acoustic sensor networks: research challenges. *Ad Hoc Networks*, 3(3):257–279, 2005.
- [3] Archana Toky, Rishi Pal Singh, and Sanjoy Das. Localization schemes for underwater acoustic sensor networks - a review. *Computer Science Review*, 37:100241, 2020.
- [4] P. Rizzo. 17 - sensing solutions for assessing and monitoring underwater systems. In M.L. Wang, J.P. Lynch, and H. Sohn, editors, *Sensor Technologies for Civil Infrastructures*, volume 56 of *Woodhead Publishing Series in Electronic and Optical Materials*, pages 525–549. Woodhead Publishing, 2014.
- [5] Imane Salhi, Martyna Poreba, Erwan Piriou, Valerie Gouet-Brunet, and Maroun Ojail. Chapter 8 - multimodal localization for embedded systems: A survey. In Michael Ying Yang, Bodo Rosenhahn, and Vittorio Murino, editors, *Multimodal Scene Understanding*, pages 199–278. Academic Press, 2019.
- [6] B Mishachandar and S Vairamuthu. An underwater cognitive acoustic network strategy for efficient spectrum utilization. *Applied Acoustics*, 175:107861, 2021.
- [7] David Munoz, Frantz Bouchereau, Cesar Vargas, and Rogerio Enriquez. Chapter 1 - the position location problem. In David Munoz, Frantz Bouchereau, Cesar Vargas, and Rogerio Enriquez, editors, *Position Location Techniques and Applications*, pages 1–22. Academic Press, Oxford, 2009.
- [8] L. Bjorno. Chapter 14 - underwater acoustic measurements and their applications. In Thomas H. Neighbors and David Bradley, editors, *Applied Underwater Acoustics*, pages 889–947. Elsevier, 2017.
- [9] N. Crasta, D. Moreno-Salinas, A.M. Pascoal, and J. Aranda. Multiple autonomous surface vehicle motion planning for cooperative range-based underwater target localization. *Annual Reviews in Control*, 46:326–342, 2018.
- [10] S. Longhi, A. Monteriù, and M. Vaccarini. Cooperative control of underwater glider fleets by fault tolerant decentralized mpc. *IFAC Proceedings Volumes*, 41(2):16021–16026, 2008. 17th IFAC World Congress.
- [11] Jian Lu, Xu Chen, Maoxin Luo, and Yanran Zhou. Cooperative localization for multiple auvs based on the rough estimation of the measurements. *Applied Soft Computing*, 91:106197, 2020.
- [12] Shaonan Li, Wenyu Qu, Chunfeng Liu, Tie Qiu, and Zhao Zhao. Survey on high reliability wireless communication for underwater sensor networks. *Journal of Network and Computer Applications*, 148:102446, 2019.
- [13] Lin Ma, T. Aaron Gulliver, Anbang Zhao, Chunsha Ge, and Xuejie Bi. Underwater broadband source detection using an acoustic vector sensor with an adaptive passive matched filter. *Applied Acoustics*, 148:162–174, 2019.
- [14] Shuxia Huang, Shiliang Fang, and Ning Han. Iterative matching-based parameter estimation for time-scale underwater acoustic multipath echo. *Applied Acoustics*, 159:107094, 2020.
- [15] Daniela Mercedes Martínez Plata and Ángel Gabriel Andrade Reátiga. Evaluation of energy detection for spectrum sensing based on the dynamic selection of detection-threshold. *Procedia Engineering*, 35:135–143, 2012. International Meeting of Electrical Engineering Research 2012.
- [16] B. Sarala, S. Rukmani Devi, and J. Joselin Jeya Sheela. Spectrum energy detection in cognitive radio networks based on a novel adaptive threshold energy detection method. *Computer Communications*, 152:1–7, 2020.
- [17] L. KOPP. Détection et estimation en traitement d’antenne : théorie. *Techniques de l’ingénieur Systèmes radars*, TIB591DUO(te5225), 2003.
- [18] L. Kopp, G. Bienvenu, and M. Aiach. New approach to source detection in passive listening. In *ICASSP ’82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 779–782, 1982.
- [19] Marco Gori. Chapter 5 - deep architectures. In Marco Gori, editor, *Machine Learning*, pages 236–338. Morgan Kaufmann, 2018.
- [20] Dogan Ibrahim. Chapter 5 - pic18 microcontroller development tools. In Dogan Ibrahim, editor, *SD Card Projects Using the PIC Microcontroller*, pages 257–297. Newnes, Boston, 2010.
- [21] Robert Oshana. 5 - dsp architectures. In Robert Oshana, editor, *DSP Software Development Techniques for Embedded and Real-Time Systems*, Embedded Technology, pages 123–158. Newnes, Burlington, 2006.
- [22] John H. Davies. Chapter 2 - the texas instruments msp430. In John H. Davies, editor, *MSP430 Microcontroller Basics*, pages 21–42. Newnes, Burlington, 2008.
- [23] Redouane Es-sadaoui, Lahoucine Azergui, Youssef Ghanam, and Jamal Khallaayoune. Design and experimentation of a low-power iot embedded system for wireless underwater sensing. In *International Conference on Wireless Networks and Mobile Communications, WINCOM 2017, Rabat, Morocco, November 1-4, 2017*, pages 1–6. IEEE, 2017.
- [24] Jinwang Yi, Diba Mirza, Ryan Kastner, Curt Schurgers, Paul Roberts, and Jules Jaffe. Toa-ts: Time of arrival based joint time synchronization and tracking for mobile underwater systems. *Ad Hoc Networks*, 34:211–223, 2015. ADVANCES IN UNDERWATER COMMUNICATIONS AND NETWORKS.
- [25] W. Kenneth Jenkins, Douglas L. Jones, and Bill J. Hunsinger. 28 - discrete-time signal processing. In Wendy M. Middleton and Mac E. Van Valkenburg, editors, *Reference Data for Engineers (Ninth Edition)*, pages 28–1–28–39. Newnes, Woburn, ninth edition edition, 2002.
- [26] Marc T. Thompson. Chapter 2 - review of signal processing basics. In Marc T. Thompson, editor, *Intuitive Analog Circuit Design (Second Edition)*, pages 15–52. Newnes, Boston, second edition edition, 2014.
- [27] M.A. Do. Modern methods of improving the range accuracy of ctfm sonars. *Ultrasonics*, 22(3):110–114, 1984.
- [28] Ibrahim Abdulhadi Sulaiman, Hussain Mohammad Hassan, Mohammad Danish, Munendra Singh, P.K. Singh, and Manisha Rajoriya. Design, comparison and analysis of low pass fir filter using window techniques method. *Materials Today: Proceedings*, 2020.

A New Corner Detection Operator for Multi-Spectral Images

Hassan El Houari¹
LaGuardia Community College,
CUNY, New York, USA

Ahmed Fouad El Ouafdi²
Ibn Zohr university,
Agadir, Morocco

Abstract—Corner detection is a crucial image processing technique that has a wide range of application, including motion detection, image registration, video tracking, and object recognition. Most proposed approaches for corner detection are based on gray-scale images, despite it has been shown that color information can greatly improve the quality of corners detection. This paper aims to introduce a new operator that identifies the second-order image information for multi-spectral images. The operator is developed using the multi-spectral gradient and differential structures of the image. Consequently, the eigenvectors of the proposed operator are used for detecting corners. A comparative study is conducted using synthetic and real images, and the result confirms that the proposed approach performs better compared with two other approaches for detecting corners.

Keywords—Corner detection; multi-spectral; operator

I. INTRODUCTION

Corner points are considered as important structural elements for extracting features of local images. The word corner is commonly referred to as a point of interest in the image with abruptly changing intensity and/or contours in all directions at the same time. Detection of such points are popular in a wide range of applications, such as motion tracking, images matching, robot navigation, object detection and recognition and image registration [1], [2], [3], [4], [7], [9], [8]. Although the corners can clearly be recognized by the human vision system, the automated detection of the exact corner location is a non-trivial task. A good corner detector must fulfill a number of eligible criteria, i.e., discern between real and false corners, reliably identify corner positions, be robust in terms of noise and efficiency.

Numerous forms of corner detection have been published in literature during the past few decades. [7], [10], [6], [5], [11], [12]. Majority of these methods can be classified into two categories depending on whether the method is contour- or intensity-based. In what follows, a review of some corner approaches from both categories. Dating back to 1977, Moravec [13] proposed one of the early corner detection algorithms, in which the corner point is described as a point of low self-similarity. The Moravec's idea was developed to propose the Harris' algorithm by using the first order derivatives to approximate the second derivatives [14]. Later, the operator used in Harris' algorithm was extended to space-time [15]. Mikolajczyk and Schmid [16] proposed a corner detector based on the Harris corner detector and the Gaussian scale space representation. The Harris's algorithm was developed for three-dimensional multi-spectral images based on correlation [17]. A multi-scale point detector based on the Gabor Wavelet

principle is presented [18]. Recently, an adaptive corner detection method based on deep learning is proposed [19].

Most of these approaches are built on the assumption that the corners correspond to abrupt changes in contour directions, and are based on this observation by examining the first and second derivative of the image to locate the corners. In case of mono-spectral images (gray level), the first derivative is roughly computed by the gradient vector and the second derivative by the Hessian matrix. In case of multi-spectral images, the images are considered as dimension two differential manifold. Thus, multispectral contours are identified by the eigenvalues and eigenvectors of the metric tensor estimated by the product of the transposed Jacobian matrix with itself [20]. For the extraction of multi-spectral information of second order, the basic approach is first to separately calculate the Hessian matrix of each band, then realize a direct Hessian matrix sum to produce the final second order differential matrix [21]. However, by performing a direct sum, the terms of the Hessian matrix may be opposite signs, so that the amount may lead to the cancellation of the second derivatives. To solve the problem, a quaternion-based method formed by the Hessian matrix of each band was proposed [22]. However, this method has proved time consuming, because the calculation of the eigenvalues requires the singular values of the quaternion decomposition.

To the best of our knowledge, an operator that detects a multi-spectral image's second-order information has not yet been proposed. From the excesses of the multi-spectral gradient and the differential structure of the image, a multi-spectral operator to identify the second-order information of color images is developed initially, then, the eigenvalues of this operator is used for detecting corners.

II. NOTATION AND PRELIMINARIES

Given an m -bands image defined by $I : \mathbb{R}^2 \rightarrow \mathbb{R}^m$, that maps a point (x, y) in the image plane to an vector $I(x, y) = (I_1(x, y), \dots, I_m(x, y))$ in \mathbb{R}^m . The variations of an image are evaluated by the change in the image values in an infinitesimal displacement. This could be represented by the differential

$$dI = \sum_{k=0}^m \frac{\partial I_k}{\partial x_k}$$

The squared norm of dI , which indicates how much the image value varies in any direction, is given as

$$dI^2 = \sum_{k=0}^m \sum_{h=0}^m \frac{\partial I_k}{\partial x_k} \frac{\partial I_h}{\partial x_k}$$

Using tensor notation, Di Zenzo [20] introduced the first definition of the gradient of a multi-spectral image. Let J be the jacobian matrix of I . Then, the metric tensor is approximated by the matrix

$$J^T J = \begin{pmatrix} E & F \\ F & G \end{pmatrix}, \quad (1)$$

Where

$$E = \sum_{k=0}^m \left(\frac{\partial I_k}{\partial x} \right)^2,$$

$$F = \sum_{k=0}^m \frac{\partial I_k}{\partial x} \frac{\partial I_k}{\partial y},$$

$$G = \sum_{k=0}^m \left(\frac{\partial I_k}{\partial y} \right)^2$$

At each point on the image, there are two main quantities to be know; the direction along which the function I has the maximum rate of change, and the absolute value of this maximum rate of change. The variations of multispectral image I are extreme in the directions of the eigenvectors of the matrix tensor $J^T J$. If the maximum and minimum contrasts (corresponding to the largest and smallest eigenvalues of $J^T J$) are represented by λ_{max} and λ_{min} , respectively, then, such extreme values can be calculated as

$$\lambda_{max} = \frac{E + G + \sqrt{(E - G)^2 + 4F^2}}{2}$$

$$\lambda_{min} = \frac{E + G - \sqrt{(E - G)^2 + 4F^2}}{2}$$

Let N be the associated eigenvector to λ_{max} . This vector corresponds to the direction of the maximum variation and it is given by

$$N(x, y) = (\cos(\theta), \sin(\theta)),$$

with

$$\theta = \frac{1}{2} \arctan \left(\frac{2F}{E - G} \right) + n\pi \quad \text{for } n \in \mathbb{Z}.$$

If the multi-spectral image as a surface in the space denoted by $\mathcal{I}_I = \{I(x, y), (x, y) \in \mathbb{R}^2\}$. The image is, thereby, a two-dimensional surface embedded in \mathbb{R}^m . At each point $p(x, y) \in \mathcal{I}_I$, let $\mathcal{T}_p I$ denotes the tangent plane of the surface \mathcal{I}_I generated by the two vectors $\vec{U} = \frac{\partial I}{\partial x}$ et $\vec{V} = \frac{\partial I}{\partial y}$. Therefore, the vector N may be represented in the basis (\vec{U}, \vec{V}) in the form:

$$N(x, y) = \cos(\theta)\vec{U} + \sin(\theta)\vec{V}. \quad (2)$$

as shown in figure 1.

III. DESCRIPTION OF THE NEW MULTI-SPECTRAL OPERATOR

In the case of multi-spectral image, the components $I_1 = R$, $I_2 = G$ and $I_3 = B$ are identified. As noted above, the vector N points to the direction of the maximum change of the multi-spectral contour and orientation angle depends of the coordinate (x, y) . Given that the corners are localized on the points that correspond to an abrupt change in the orientation of the contour, at first step, the Jacobian matrix J_N of N with variables x and y is calculated. Then the eigenvalues of the matrix $J_N^T J_N$ are incorporated in a decision rule to locate the corners. Taking into account the fact that the two vectors I_x and I_y which generate the tangent plane $\mathcal{T}_p I$ change from one point to another as shown in Figure 1.

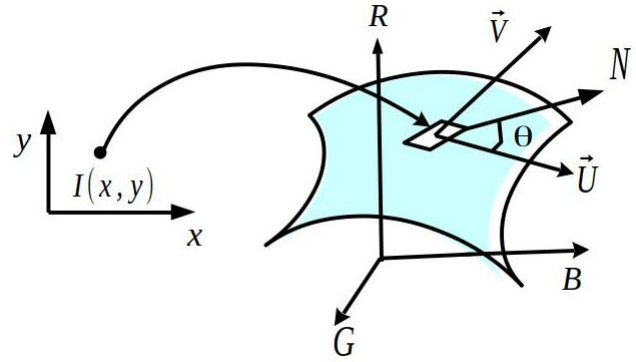


Fig. 1. Parametrization of Image in Color Space.

Applying the chain rule to (2), the Jacobian matrix J_N of the vector N at point (x, y) can be formulated as 2×3 matrix in term of the two vectors J_1 and J_2 as follows

$$J_N = [J_1 \ J_2] \quad (3)$$

where

$$J_1 = \frac{\partial N}{\partial x} = \begin{bmatrix} \cos(\theta)(R_{xx} + \theta_x R_y) + \sin(\theta)(R_{yx} - \theta_x R_x) \\ \cos(\theta)(G_{xx} + \theta_x G_y) + \sin(\theta)(G_{yx} - \theta_x G_x) \\ \cos(\theta)(B_{xx} + \theta_x B_y) + \sin(\theta)(B_{yx} - \theta_x B_x) \end{bmatrix}$$

and

$$J_2 = \frac{\partial N}{\partial y} = \begin{bmatrix} \sin(\theta)(R_{yy} - \theta_y R_x) + \cos(\theta)(R_{xy} + \theta_y R_y) \\ \sin(\theta)(G_{yy} - \theta_y G_x) + \cos(\theta)(G_{xy} + \theta_y G_y) \\ \sin(\theta)(B_{yy} - \theta_y B_x) + \cos(\theta)(B_{xy} + \theta_y B_y) \end{bmatrix}$$

The indices correspond to the first and second derivatives with respect to x and y . θ_x and θ_y are the partial derivatives of the angle θ given by

$$\theta_\delta = \frac{F_\delta(E - G) - F(E_\delta - G_\delta)}{(E - G)^2 + 4F^2} \quad \text{pour } \delta = x, y.$$

Here the second-order derivatives of $R_{\bullet\bullet}$, $G_{\bullet\bullet}$ and $B_{\bullet\bullet}$ are computed by convolution of the color channels R , G and B with a second order Gaussian derivative mask. The powerful Gaussian property guarantees the existence and continuity of the second derivatives of I .

In the Jacobian matrix (3), the presence of terms of second derivatives of the three bands of the image, not as a direct sum, but instead as combination of rotation angles and their derivatives, as well as the terms of the first derivative of the image. Now, to quantify the change in the vector N , consider the matrix $J_n^T J_N$ that can be interpreted as the matrix that approximates the metric tensor of the space formed by the vectors N . As in the case of the eigenvalues of matrix (1) that determine the multi-spectral contour, the eigenvalues λ_{max} and λ_{min} of the matrix $J_N^T J_N$ quantify variations on this contour, which allows to identify the corners which are characterized by an abrupt change of the orientation of the contour. As rule decision to detect the corners, the following function [14] is used:

$$R_\kappa = \lambda_{max}\lambda_{min} - \kappa(\lambda_{max} + \lambda_{min})^2 \quad (4)$$

$$= \det(J_N^T J_N) - \kappa \text{trace}^2(J_N^T J_N),$$

Where κ is the sensitivity setting. The smaller the value of κ , the more likely it is to detect the corners with acute angles.

IV. EXPERIMENTAL AND DISCUSSIONS

In this section, the results of the proposed corner detector on two different applications are presented, the first application is to detect corners in an image, and the second application is to track corners in a video sequence.

A. Localization of Corners.

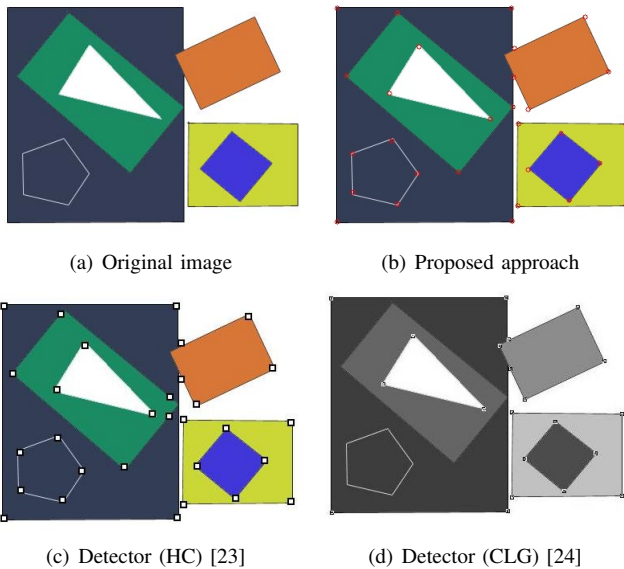


Fig. 2. Detection Results of the Corners on a Synthesis Image.

In this section, a comparison of the proposed method of corner detection with two standard methods in literature is reported; the first one is an extension of the harris's detector [14] to color images (HC) proposed in [23], this method is based on approximation of the auto-correlation of the gradient in different directions. The second approach is based on local and global curvatures (CLG) for the detection of corners on gray-scale images [24].

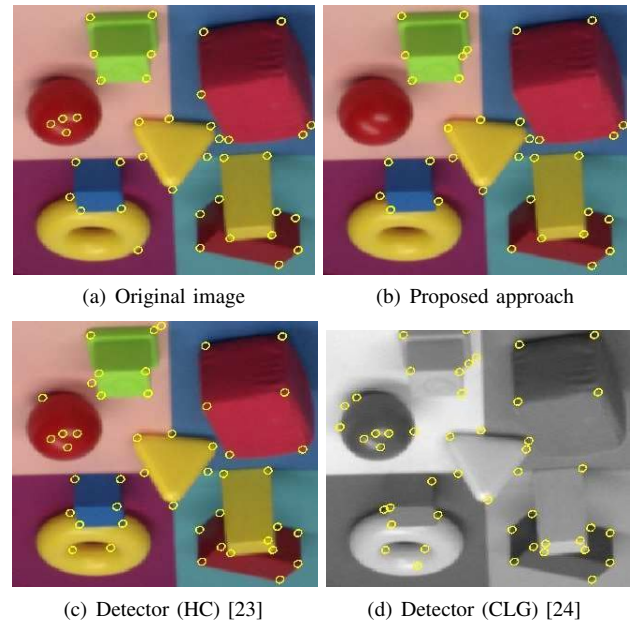


Fig. 3. Detection Results of the Corners on a Synthesis Image.

A comparative study of different corner detection methods [25] show that these two approaches demonstrate good performance compared to other approaches in the literature. For the calculation of derivatives in expressions (1) (3) and (III), a Gaussian bypass filters of first and second orders are applied, the best results are obtained when the sensitivity setting (4) is attached to the value 0.05.

B. Qualitative Comparison

In Figure2 and figure 3, the results of three approaches on two synthetic images are reported, while in figure4 and figure 5 the comparative corner detection methods are applied on three real images. Note that the detection of corners is almost perfect in computer graphics for the proposed approach and the comparative detection methods based on the color (HC), however, the method (CLG) could not properly detect many corners in the image, especially, those of the white hexagon.

The difference between the three methods is more visible when applied to real image in Figure 4. The method (CLG) furthermore, while detecting a majority of corners, it also detects many false corners including the texture of plants and grass, as for the approach (HC), it could not detect a few important corners in the image, especially those formed by shadows and points which three or more contour regions meet. The approach proposed in this paper has led to good results of detection on real images, because most of the corners are located all in minimizing false detection.

Illumination has a strong influence on images and consequently on corners detection. Hence, images of the same scene under different illuminations can be very dissimilar, which greatly affects the detection of interest points like corners. In figure 6.a one more illumination source was added in the image on the right. There were 116 corners detected in the original image and only 118 detected in the right image, while in Figure 6.b the orientation of the illumination was changed, a difference between the four corners between the original and

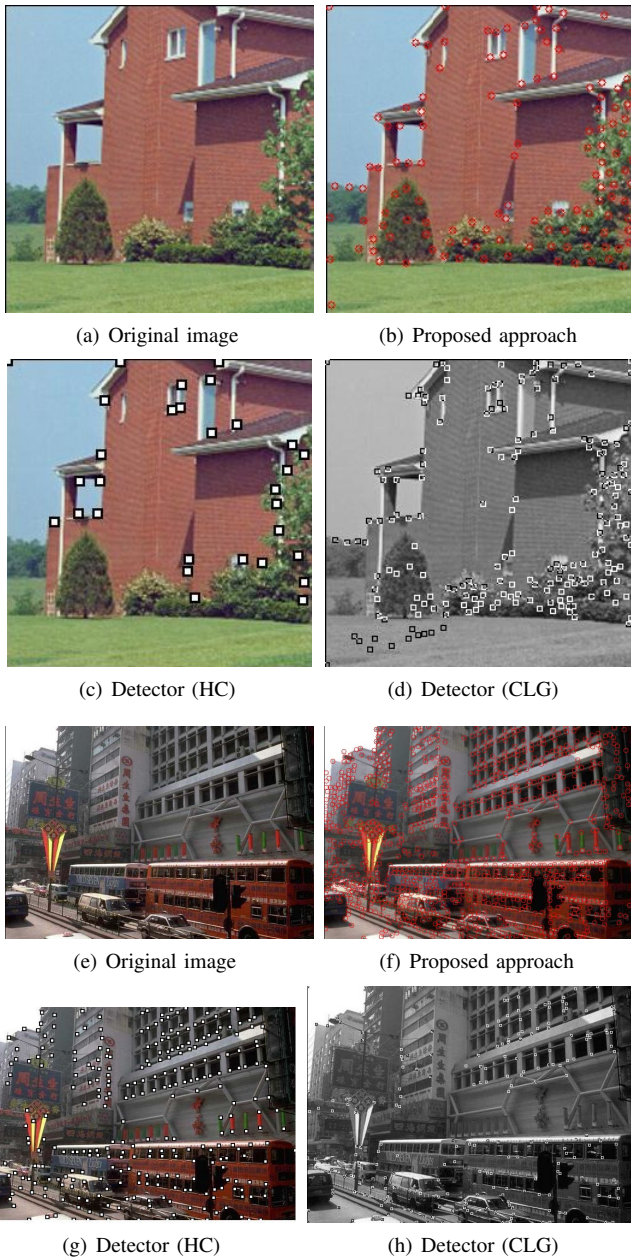


Fig. 4. Results of Detection of the Corners of the Actual Images.

modified images is noticed. Adding illumination invariance seems to have a relatively small effect in this example shown in Fig 7,8.

C. Quantitative Comparison

On noisy images, the corner detection methods are applied on a noisy artificial image and on a noisy real laboratory image. The artificial and real images illustrated in figure 9 and figure 9 are generated by adding a Gaussian noise with standard deviation $\sigma = 20$. As illustrated in 9 and figure 9, it can be observed that the proposed corner detection approach detect all corners from the noisy image. The Table 1 presents a quantitative comparison of the corner detecting methods. It can be clearly seen from the mean localization error values

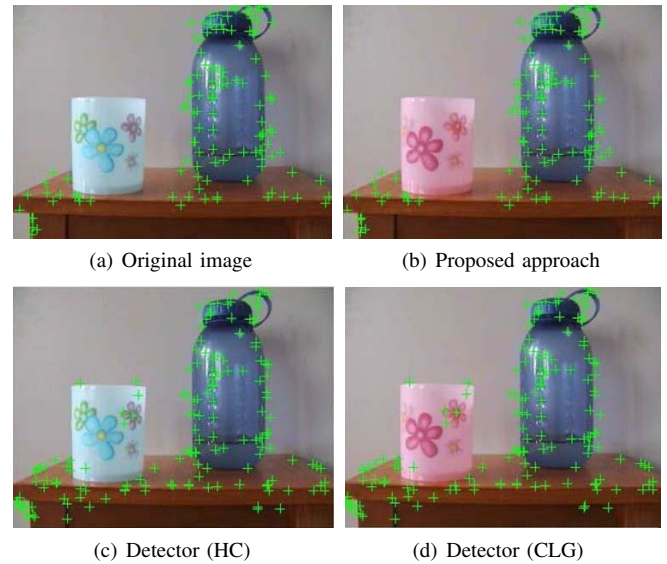


Fig. 5. Results of Detection of the Corners of the Actual Images.

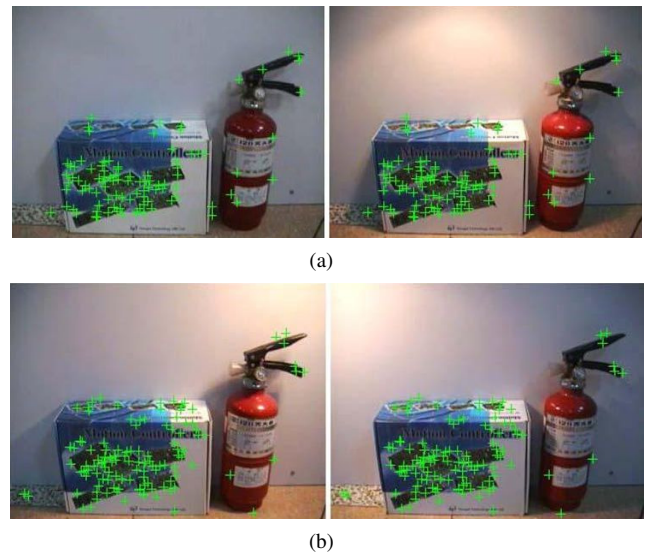


Fig. 6. Illumination and Orientation Changes.

that the proposed approach performs well in corner localization accuracy.

D. Location and Tracking of Points of Interest

Points of interest such as the corners are commonly used for tracking objects [26]. From the perspective of an automatic

TABLE I. STATISTICS ABOUT COMPARISON OF CORNER DETECTION

	Corners	Missed	Missed	Localization error
Laboratory Image				
Proposed	294	12	13	0.658
Detector (HC) [23]	287	21	14	0.954
Detector (CLG) [24]	275	33	11	1.254
Toys image				
Proposed	36	0	0	0.367
Detector (HC) [23]	35	0	0	0.528
Detector (CLG) [24]	33	5	0	1.058

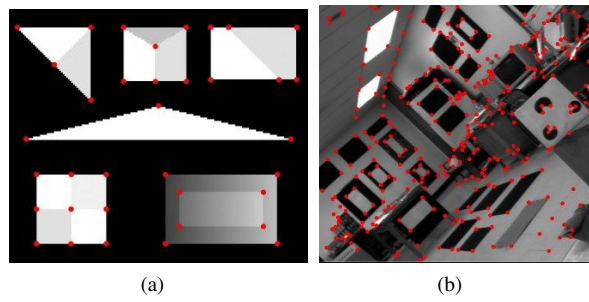


Fig. 7. Ground-truth Corners of Original Images Test.

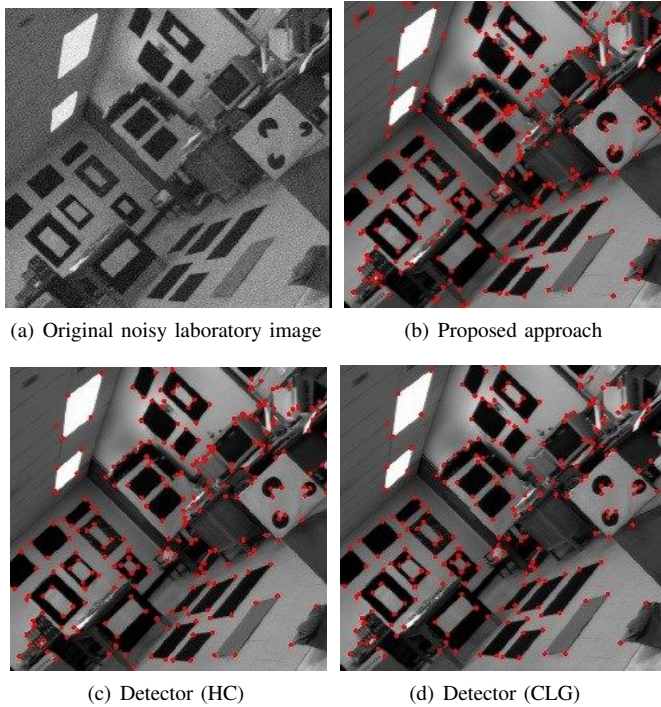


Fig. 8. Results of Corners Detection on Noisy Laboratory Image.

analysis of facial expressions, the proposed corner detector is applied in this paragraph to track the points of interest in the face in a video sequence. An automatic analysis of facial expressions system generally consists of three main phases; face detection, components extraction and finally the classification of facial expression. Extraction of facial components passes mainly through the localization and eye-tracking in the sequence video [27]. As a first step, the method of locating eyes in the face using the method proposed by Viola and Jones [28] is applied, after this initial stage, the proposed operator is used to generate corners, these feature points are used for eye tracking in the video by following the approach proposed by Shi and Tomasi [26]. Figure 10 shows an example application of the proposed operator for eye tracking. Initially in Figure 10(a), an initial location is performed for eyes detection method based on a cascade of classifiers boosted [28], then a tracking feature points is performed to estimate the position of his points in the following images, as illustrated in figures 10(a) - 10(f).

In a third application reported in Figure 11, the point cloud generated to track each eye is filtered to keep the two most

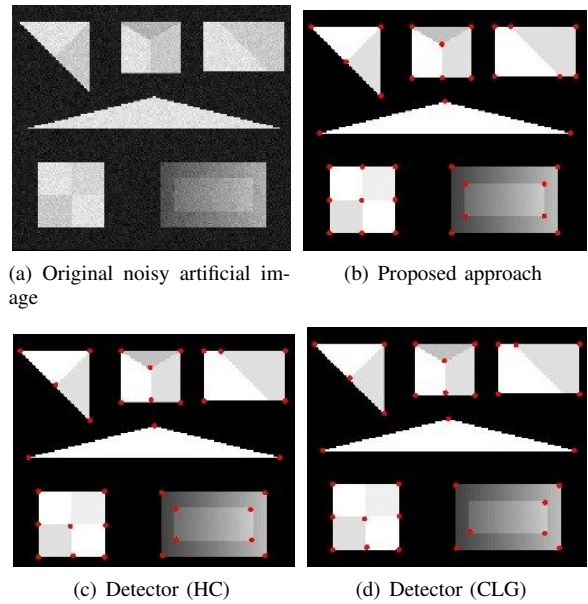


Fig. 9. Results of Corners Detection on Artificial Noisy Image.

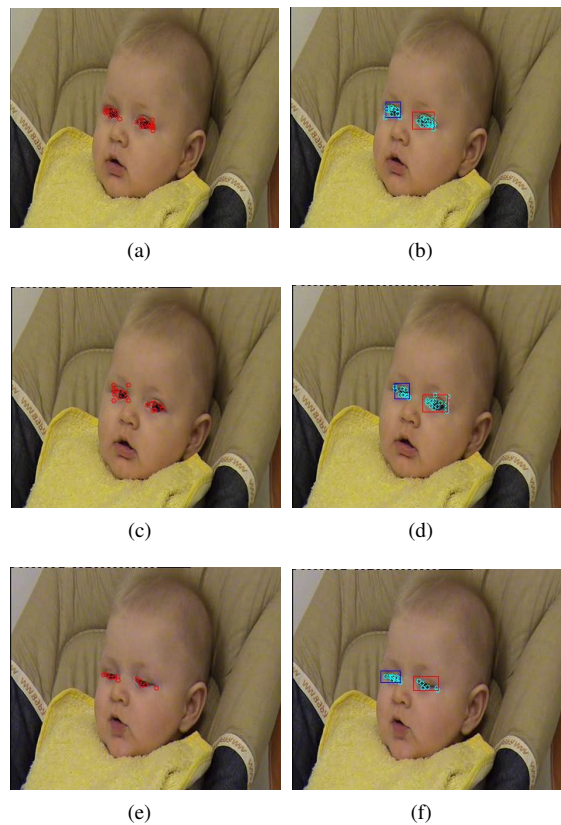


Fig. 10. Eye Tracking in a video Sequence.

important corners in both ends of the eye, these two points are essential to the realization of an automatic analysis system of facial expressions [27]).

V. CONCLUSION

In this paper, a new operator for corner detection is introduced. Initially a multi-spectral operator was developed to

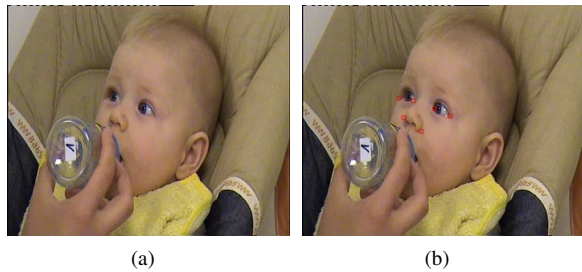


Fig. 11. Locations of Corners of the Eyes.

identify the second-order information of color images, based on its multi-spectral gradient and differential structures. As a by-product, the eigenvectors of this operator are used to detect corners. For testing the proposed approach, two methods that are considered among the most efficient methods were chosen. Experiments on synthetic and real images show a better detection of the corners by the proposed method. These preliminary results are very promising and encouraging.

REFERENCES

- [1] A. Dutta, A. Kar, and B.N. Chatleri, *A new approach to corner matching from image sequence using fuzzy similarity index*, Pattern Recognition Letters. 32(5):712-720, 2017.
- [2] S. Gauglitz, T. Hollerer, M. Turk, *Evaluation of interest point detectors and feature descriptors for visual tracking*, International Journal of Computer Vision, 94:335-360, 2011.
- [3] H. Zhang, L. Xiao and G. Xu, *A Novel Tracking Method Based on Improved FAST Corner Detection and Pyramid LK Optical Flow*, 2020 Chinese Control And Decision Conference (CCDC), 2020, pp. 1871-1876.
- [4] P. Mainli, G. Lafruit, Q. Yang, B. Geelen, L. V. Gool, R. Lauwereins, *SIFER: Scale-invariant feature detector with error resilience*, International Journal of Computer Vision, 104: 172-197, 2013.
- [5] Wang, Junqing and Zhang, Weichuan. *A Survey of Corner Detection Methods*. 2018, 10.2991/iceea-18.2018.47.
- [6] He Yarui, Li Yunhong, Fang Qiaochu *A survey of image corner detection methods*. HP3C, March 2019, Pages 123–127.
- [7] Liu, Y., Yu, H., Yang, W., Li, L. *SAR image registration using SAR-FAST corner detection*. J. Electron. Inf. Technol. 39(2), 430–436, 2017
- [8] Chengfeng Jian 1 ; Xiaoyu Xiang 1 ; Meiyu Zhang 1 *Mobile terminal gesture recognition based on improved FAST corner detection*, Volume 13, Issue 6, 10 May 2019, p. 991 – 997.
- [9] D. Wang et al., *Combined Use of FCN and Harris Corner Detection for Counting Wheat Ears in Field Conditions*, in IEEE Access, vol. 7, pp. 178930-178941, 2019.
- [10] W. Yu, G. Wang, C. Liu, Y. Li, Z. Zhang and K. Liu, *An Algorithm for Corner Detection based on Contour*, 2020 Chinese Automation Congress (CAC), 2020, pp. 114-118,
- [11] S. Chen, H. Meng, C. Zhang, C. Liu A, *A KD curvature based corner detector*. Neurocomputing 173: 434-441 (2016).
- [12] J. Chu, J. Miao, G. Zhang, L. Wang, *Edge and corner detection by color invariants*. Optics and Laser Technology, 2013 - Elsevier
- [13] H. P. Moravec., *Towards Automatic Visual Obstacle Avoidance*, Proc. 5th International Joint Conference on Artificial Intelligence, pp. 584, 1977.
- [14] C. Harris et M. Stephens, *A Combined Corner and Edge Detector* Alvey Vision Conference, 1988.
- [15] I.-Laptev and T. Lindeberg *RSpace-time interest points*, IEEE International Conference on Computer Vision. pp. 432–439, 2003.
- [16] K. Mikolajczyk, K. and C. Schmid, *Scale and affine invariant interest point detectors* International Journal of Computer Vision. 60 (1): 63–86. 2004.
- [17] Y. Li, W. Shi and A. Liu , *A Harris corner detection algorithm for multispectral images based on the correlation* Proc. 6th International Conference on Wireless, Mobile and Multi-Media (ICWMMN 2015)
- [18] W. Yussof and M. Hitam, *Invariant Gabor-based interest points detector under geometric transformation*. Digital Signal Process. 25, 190–197 (2014)
- [19] L. Wang, K. Han and H. Sun, *An Adaptive Corner Detection Method Based on Deep Learning*, Chinese Control Conference (CCC), Guangzhou, China, 2019,
- [20] S. Di Zenzo, *A note on the gradient of a multi-image*, Comput. Vision. Graph. vol.33, p.116-125, 1986.
- [21] A. Ming et H. Ma, *A blob detector in color images* Proc. 6th ACM int. conf. on Image and video retrieval, p.364-370, 2007.
- [22] L. Shi, B. Funt et G. Hamarneh, *Quaternion Color Curvature* Proc. IST Sixteenth Color Imaging Conference, Portland, 2008.
- [23] J. van de Weijer, T. Gevers et J.-M. Geusebroek, *Edge and corner detection by photometric quasi-invariants* IEEE Trans. Pattern Anal. Mach. Intell., vol.27,p.625-630, 2005.
- [24] C.H. Xiao et N.H.C. Yung, *Corner detector based on global and local curvature properties* Opt. Eng., vol. 47(5), p.057008, 2008.
- [25] J. L. A. Jakas, A. Al-Obaidi et Y. Liu. *A comparative study of different corner detection methods*. CIRA'09, p.15-18,2009.
- [26] J. Shi et C. Tomasi *Good Features to Track*, CVPR 94, p. 593-600, 1994.
- [27] Y.-Li. Tian, T. Kanade et J. Cohn *Recognizing action units for facial expression analysis*, IEEE Trans. Pattern Anal. Mach. Intell. , Vol. 23, No. 2,p. 97 - 115, February, 2001.
- [28] P. Viola et M. J. Jones. *Robust Real-Time Object Detection*, CVPR 2001, Vol. 1 p. I-511-518.

Fast Fractal Coding of MRI Images using Deep Reinforcement Learning

Bejoy Varghese¹, S. Krishnakumar²

Angmaly, Federal Institute of Science and Technology, Ernakulam, Kerala, India¹
STAS, M G University Research Centre, Ernakulam, Kerala, India²

Abstract—This paper presents an algorithm based on Fractal theory by using Iterated Function Systems (IFS). An efficient and fast coding mechanism is proposed by exploiting the self similarity nature in the Brain MRI images. The proposed algorithm utilizes Deep Reinforcement Learning (DRL) technique to learn the transformations required to recreate the original image. We avail of the Adaptive Iterated Function System (AIFS) as the encoding scheme. The proposed algorithm is trained and customised to compress the Medical images, especially Magnetic Resonance Imaging (MRI). The algorithm is tested and evaluated by using the original MR head scan test images. It learns from an existing biomedical dataset viz The Internet Brain Segmentation Repository (IBSR) to predict the new local affine transformations. The empirical analysis shows that the proposed algorithm is at least 4 times faster than the competitive methods and the decoding quality is far distinct with a reduction in the bit rate.

Keywords—Fractal compression; deep reinforcement learning; MRI image compression; deep learning; adaptive fractal coding

I. INTRODUCTION

Medical imaging has become one of the most rapidly growing fields in image processing and medical research. It includes multimodality imaging techniques like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), Elastography, and Digital Subtraction Angiography (DSA). Medical images help doctors in diagnosis, clinical staging and to prescribe therapeutics to heal the disease. But a large number of such images demand enormous storage space and the transmission bandwidth of the PACS. These requirements demand the need for high-quality medical image compression algorithms.

Efforts in reducing the encoding time with better SSIM cause the loss of information on the lesion, leading to misdiagnosis and does not achieve the required effect. Therefore, a Machine Learning-based FIC algorithm for medical images is proposed in this work. The model uses reinforcement learning techniques to gain a better compression ratio with much less encoding time. Biomedical images exhibit a very high structural similarity within the image itself. Because of this self-similarity [1], the learned algorithm can compress the image with a better compression ratio at high PSNR.

A lossless image compression seems to be more suitable for medical images, as almost all the information in it contributes a lot in the diagnosis process. The compression ratio offered by a lossless compression is very much less than the lossy compression schemes, causing it less suitable to reduce the storage and bandwidth requirements. Lossy compression schemes offer a very high level of compression ratio by omitting certain information in the source image. This causes

certain degradations in the reconstructed image, which may lead to inevitable loss of information. So many researches in this area attempt to improve the reconstruction quality of the encoded images at a fixed code rate. At the same time it's evident that an efficient lossless compression technique can play a crucial role in managing the storage space and transmission bandwidth. There are different ways to look into this problem.

The primary lossy compression technique such as JPEG [2] uses the method of identifying the information in terms of frequency components, which is more sensitive to the human eye. Another lossy compression method such as Fractal doesn't consider any of the frequency information, instead it looks for the similarities present in an image. Fractal coding is a lossy coding compression scheme, which utilizes the self similarities in an image. Fractal-based Image Compression (FIC) is rarely used in the area of biomedical images. It has been discarded due to the increased encoding time complexity of the existing algorithms.

If the lossless image compression is considered, PAQ algorithm exhibits a better performance as compared to the similar algorithms as the predictor in the PAQ coder takes a decision only based on the weighted probabilities from a large number of predictors. However, PAQ is a well stitched algorithm to compress the text rather than an image. Another image file format, TIFF is suitable to archive images, as it can hold images both in lossy and lossless schemes [3].

Hence it is observed that there are two ways to develop an efficient compression scheme for medical images. The first is focussed on the self similarity aspect of medical images, which can help in calculating the affine transformations required to reach a single fixed point. This may help us to achieve a better compression ratio and less decoding time, as promised by the fractal theory. The second way focuses on the computational theories to reduce the algorithm complexity and thus the encoding time. It is observed that a Deep Reinforcement Learning algorithm (DRL) [4] is capable of predicting the fractal similarities in an image with appreciably less time in comparison with the classical and Adaptive IFS compression schemes [5].

Iterated Function System (IFS) is a well accepted method to generate the fractals. It is a finite collection of contractive maps w_i and contractivity factors s_i in the complete metric space (X, d) , where $w_i: X \rightarrow X$ and $i = 0, 1, \dots, n$. The representation of IFS is $\{X; w_1, w_2, \dots, w_n\}$ with its contractivity factor $s = \max\{s_1, s_2, \dots, s_n\}$. The value of s lies in between $0 & 1$; $1 > s \geq 0$. By applying the contractive maps in a recursive manner on any arbitrary values leads to the generation of a fixed single point,

which is called the attractor of a particular IFS. The thrust process in fractal generation is to generate the contractive maps or affine transformations. The most celebrated Collage theorem [6] helps to identify the affine transformations that minimizes the distance between the given subsets. So the required process is to compute the coefficients of contractive maps and its probability factors. This can be calculated by using the Markov operator, which itself is a contractive map in the complete metric space of probabilities. It has been found that the collage theorem is the most suitable method to calculate the IFS with probabilities, hence to identify the fixed point of the block under consideration. The modified version of classical fractal compression method based on probability and multiscaling division proves to be more efficient in terms of encoding time and computational complexity [7].

In the case of biomedical images, the number of similar patterns are immense and repetitive over the time scale. Majority of the biomedical imaging techniques produce multi-modality images and the resemblance in the image of the same organ for different patients is high [8]. Hence, the initial idea is to develop an adaptive and efficient Fractal based method, capable of using the self-similarity essence in the human body to compress the respective biomedical images.

The paper is structured as follows. Section 2 explains the fundamentals of fractal compression and latest findings in the field. The idea of reinforcement learning and its applications to fractal compression is explained in section 3. Section 4 explains the proposed method. Results from the work are discussed in section 5. Section 6 concludes the paper.

II. RELATED WORKS

The idea of applying IFS in fractal image compression is suggested by M. Barnsley [9]. Later on his student A Jacquin [10], [11] could automate the basic IFS by using Markov contractive operators. These developments upsurge the use of IFS based fractal image compression in 1990. But still the speed of the encoding process did not meet the practicality. Y Fisher [12] developed a quad tree based partitioning system that did much to ameliorate the encoding process.

Despite all these advancements, the biggest problem still pulls back the fractal based image compression is the calculation of affine transformations, by comparing domain and range blocks. B. Hurtgen [13] suggested to consider the average pixel intensities and block variances to limit the number of possible domain range comparisons. Similarly, the idea of reducing range domain comparisons by applying the Nearest Neighbour Search algorithm was suggested by D. Saupe [14]. FASON algorithm developed by Tan [15], insists to do the comparison directly without storing the domain block pool. This was followed by the use of entropy which is capable of representing the statistical characteristics of pixel data. This method is developed by Yusong Tan and is observed as the most successful method in the pile of various classical IFS algorithms. All these works were mainly oriented towards the Search-less technique in calculating the contractive maps. Wang et. al [16] developed a No-search algorithm to improve the quality of decoded images generated by the quad tree based no-search algorithm implemented by Furoo and Hasegawa's [17]. However, both the search-less and no-search algorithms could not guarantee the quality of decoded images.

All these FIC based algorithms are categorized into three: Classification based, Feature vector based methods and meta heuristic approach. The first method uses a common characteristic metric to classify the domain and range blocks into a predetermined pool. This helps to restrict the search within a limited or same class of blocks [18], [13]. But the second method needs to calculate a particular feature of the partitioned image to classify into different block pools or to discard it from a particular range- domain comparisons [14], [19]. First generation of the third method utilized Genetic Algorithms(GA) [20], Particle swarm optimization [21], ant colony which seems inefficient in terms of encoding time complexity. Because the GAs use brute force search between pairs after considering the mutation to obtain best pairs. The search may never converge, if the system can't find the best suitable domain-range pairs. This process is very similar to the exhaustive search algorithm used in baseline FIC. GA uses boundary conditions to limit the search space and hence to finish the exhaustive searching. This limitation can be overcome by utilizing the characteristics of parallel computing. On the other hand, GA is capable of offering better compression ratio and compression accuracy, which in turn leads to a better PSNR and makes it more suitable for low bit rate image compression applications [22].

Second generation meta heuristic approach adopts the statistical learning theories and proves that the computational complexity is barely minimum as it utilizes algorithms like stochastic gradient descent, the Least square optimization etc. A typical Neural Network (NN) algorithm encodes the input images to vectors in latent space and hence to make it more compact. The two main categories of NN are one-time feed forward frameworks and multi-stage recurrent frameworks [23]. Both these frameworks have their own characteristics, pros and cons. It is observed that feed forward neural networks take less time to encode and decode, as the network needs to execute fewer times. The Training phase is also easier in feed forward because the back propagation path is shorter and shallower in contrast with the recurrent networks.

Todeciri et. al [24] developed a recurrent convolution LSTM [25] based network for learned image compression and proves as an efficient method to handle the variable bit rate algorithms. There have been many variants of the basic Divisive Normalization method proposed by Balle et. al [23] and all such algorithms exhibit an excellent performance in compressing the images. Nakanishi [26] developed a 3D convolutional neural network for learning the conditional probability model to compress the image. Rippel et. al [27] proposed a model based on adversarial loss function and its decoding process is improved by Tschannen et. al [28]. He suggested using the Generative adversarial network instead of the loss function, proves to be the more suitable method to improve the decoding image quality with very low bit rate. The basic concept of policy based reinforcement learning approaches seems to be very efficient in predicting the required transformations [29], [28].

Chen et.al [30] proposed a Non-Local Attention optimization and Improved Context modeling-based image compression (NLAIC) algorithm that relies on trained deep neural networks to achieve improved rate distortion. Ma et. al [31] customized an architecture based on neural network and wavelet trans-

form capable to support both lossy and lossless compression schemes. Cheng et.al developed a flexible entropy model based on discretized Gaussian mixture likelihoods by taking the advantage of recent attention modules and is proved its efficiency in reducing the latency [32].

III. BACKGROUND

Reinforcement learning is the method of learning the best action to optimize the solution based on the reward or punishment. The system that works to learn the action is called an agent and the system that provides the reward or punishment is called environment. So to decide on the action that affects the environment, there are many algorithms that exist. One among the widely known methods is called Q learning. Q learning will decide the action based on a table called Q-Table. The algorithm continuously updates the table based on the observations and rewards from the environment. It uses the Bellman equation to calculate the action values. Bellman equation refers to a set of equations that calculate the value from the reward and discounted future values.

$$Q(s, a) = r(s, a) + \gamma \times \max_a Q(s', a) \quad (1)$$

In the Equation (1), $r(s, a)$ is the immediate reward by taking action a and in state s . $Q(s', a)$ represents the Q -value possible from the next state s' . The γ the discount factor to diminish the effects of future Q -value. This is a recursive equation, that starts with random values of Q in the initial state. On each iteration of the algorithm, the values will be updated based on the reward. In a practical implementation, The Equation (1) can be updates to include the learning rate α .

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha [R_{t+1} + \gamma \times \max_a Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (2)$$

But the major issue with Q Learning is that the size of the table becomes very large depending on the dimensionality of the input. So applying the Q Learning algorithm to relatively large data input such as images to take the decision can be a cause of slower execution. Following section presents a solution to the issue by applying a Deep Neural network to make a policy decision.

A. Deep Reinforcement Learning (DRL)

In many cases the decision process is a high dimensional problem based on the input. So to make use of the reinforcement learning technique in high dimensional space the decision process is modified by introducing a multi layer neural network for learning the policy. Depending on the size of the problem the neural network may use the deep learning technique to achieve a good accuracy. In the case of a fractal compression method, the action space or the set of transformations required to compress the image is too large to be completely known to the system. The neural network can approximate the policy function that can be used to map the states to action values. In classical RL algorithms, mapping is based on the lookup table which stores all possible combinations of state value pairs.

The input observation and reward will be used to compose a state matrix and feed to a policy network. Policy network is a deep neural network that was trained to predict the action

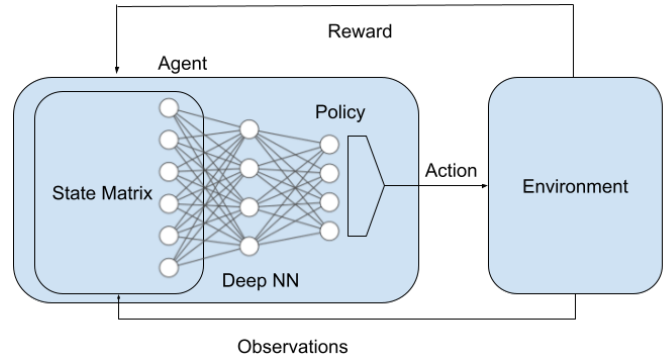


Fig. 1. Deep Reinforcement Learning System.

based on a set of states. Figure 1 shows the architecture of the deep reinforcement learning method. A multi layer neural network will predict the Q values and actions based on the state matrix. Back propagation algorithm is used to train the policy network. A multi layer neural network will predict the Q values and actions based on the state matrix. Back propagation algorithm is used to train the policy network. Deep learning method is used in conjunction with Iterated system function to achieve the fractal compression on raw image data.

IV. PROPOSED METHOD

To reduce the time complexity of the classical fractal compression methods, the brute force search is replaced by the Deep Reinforcement Learning system. Consider T represents the set of Transformation generated by the classical fractal compression for a raw image X . Then the T' represents the set transformation predicted by the DRL system. The relation between T and T' is calculated using Pearson Correlation Coefficient (PCC). PCC is -1 if the transformations are entirely different, 1 if the Transformations are equal and 0 if there is no linear correlation. Then the reward is calculated using the Equation (3).

$$r = \text{conv}(T, T') / (\sigma T \sigma T') \quad (3)$$

In the Equation (3) conv is covariance, σT is the standard deviation of T and $\sigma T'$ is the standard deviation of T' .

A. DRL Training Process

To adapt the fractal compression logic proposed system uses a function approximation such as a neural network with parameter θ to estimate the Q -Values.

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s' \sim \rho(\cdot)} [(y_i - Q(s, a; \theta_i))^2] \quad (4)$$

In Equation (4), $y_i = r + \gamma \max_{a'} Q(s', a'; \theta_{i-1})$ Where y_i is called the temporal difference and $y_i - Q$ is called temporal difference error. The ρ is called the behavioral distribution by considering the states s, a, r, s' .

To train the network to predict the transformations, the output from a classical fractal compression is considered as the true transformation value for the raw image X . The training process of DRL is described in the Code Snippet 1.

```
1 Data: Training Image set
2 Result: Trained network
3 image index=1;
4 while image index i~= size of training
  set do
5   read image;
6   range block set=partition(image);
7   domain block set=partition(image);
8   range index=1;
9   while range index = size of range
    blocks do
10    read range block;
11    domain book= transformations (
      domain blocks);
12    if Does any member of domain
      book matches with range
      block then
13      record the
        transformation;
14    end
15    else
16      record the best possible
        transformation T;
17    end
18    increase range index;
19  end
20  T' = RL agent prediction (image);
21  Calculate reward r from T and T';
22  forward the reward r to the agent.
23  end
```

Code Snippet 1: DRL Training algorithm

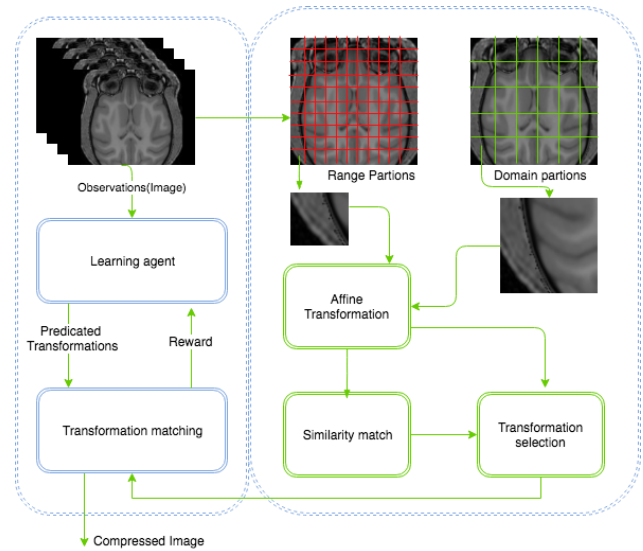


Fig. 2. System Architecture of DRL based Fractal Compression.

Figure 2 shows the system architecture of DRL based fractal compression. Raw image set is compressed using the classical fractal compression algorithm to generate the data for training the DRL. Classical compression algorithm uses the domain-range comparison to obtain the transformation set T. To avoid the time taken by the classical fractal compression, the entire raw data images are compressed in advance before training the DRL system.

The proposed system detaches the classical compression scheme from generating the transformations, once the RL algorithm learns enough to make more accurate predictions for the particular raw image inputs. To compress a new image, the raw data is passed to the DRL system to predict the transformations. The trained DRL predicts the transformation in the context of training data, in this case the MRI images. To compress any other type of images, the DRL system. The results obtained from the proposed solution is discussed in the following section.

B. Preprocessing and Model Architecture of Policy Network

The input raw image consists of three channels that follow the RGB standard. Each value in the image is 8bit size. The resolution of the image is varying from 400x400 pixels to 5120x5120 pixels. Images are re-scaled to the resolution of 800x800 pixels. The size to the input image array to the policy network is set to the maximum size of the image input. If the resolution is less than the input size of the network, the

image is aligned to the center of the 800x800 pixels and the rest of the input taken as 0. Each layer in the input image is treated separately and the transformations are combined after processing the three channels. The first layer of the neural network consists of 64000 neurons. The second, third and fourth layers include 1200,1200 and 800 neurons respectively. So the output layer of the network consists of 800 neurons, corresponding to the number of transformations. All together the network comprises 79 million parameters.

V. RESULT AND DISCUSSION

In this section we discuss some of the conducted experiments, the experimental setup and the analysis. The experiments are using the images from The Internet Brain Segmentation Repository (IBSR).

A. Experiments

We have performed 10 experiments using 3 different types of MRI images — Sagittal, Coronal and Cross-sectional.

In each experiment the training sample sizes are 500, 1500, 2500, 3500, 4500, 5500, 6500, 7500, 8500, 9500 respectively. Each experiment includes a validation set to verify the performance of the system. The size of the validation set is 20% of the training set. All the experiments use the same policy network architecture, learning algorithm and hyperparameters settings. The training of the network uses RMSProp algorithm with a mini batch size of 16 samples. The performance of the system is measured on the validation test after each experiment, by measuring the PSNR of the uncompressed image.

B. Evaluation Metrics

- 1) **Execution time:** It indicates the total time required to complete the execution of the algorithm. It does not include the time required to read or write images.
- 2) **Peak Signal to Noise Ratio (PSNR) :** Indicates the quality of the reconstructed image. It is the ratio of

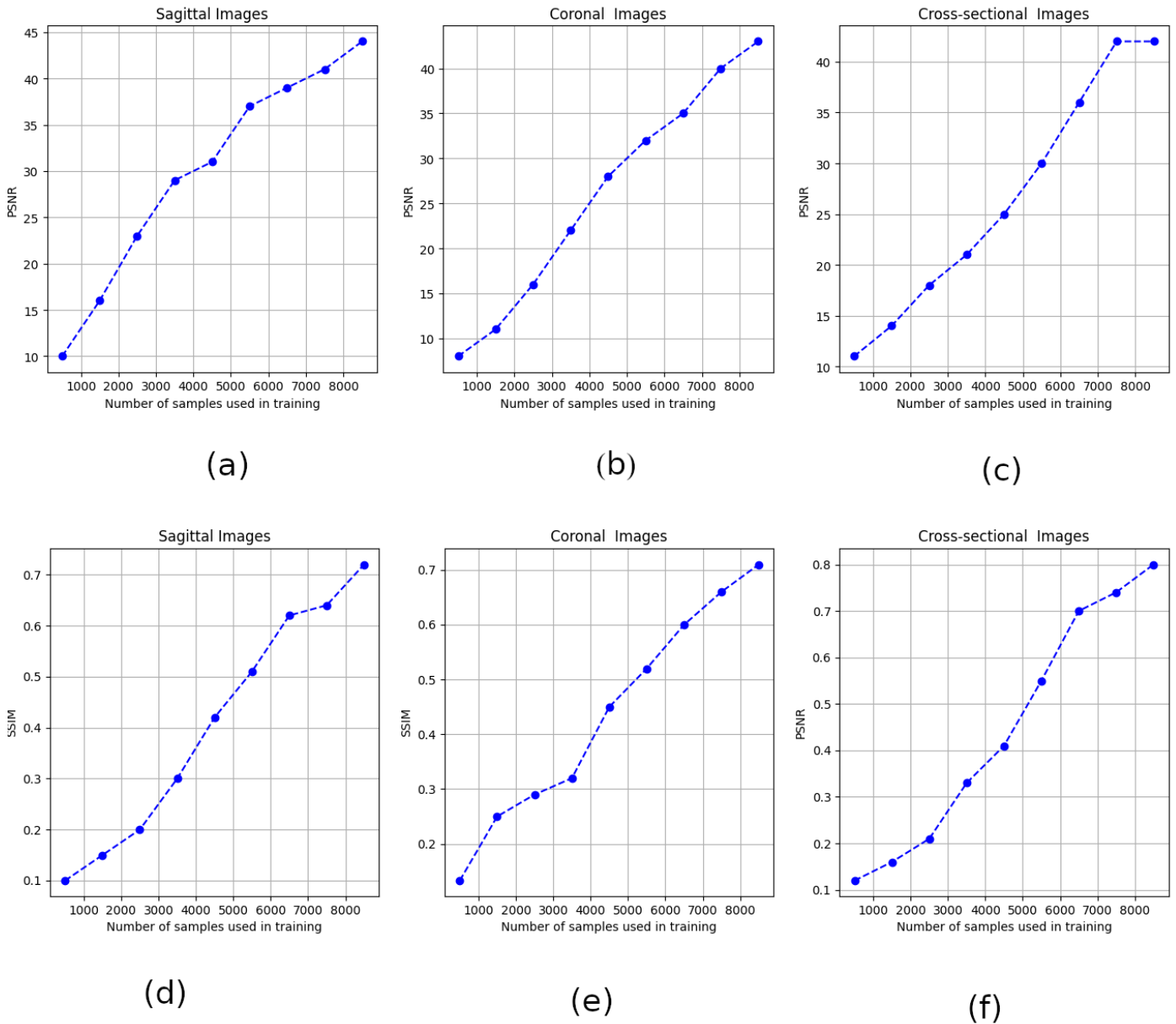


Fig. 3. (a)-(c) PSNR Variations with Respect to the Training Samples Size. (d)-(e) SSIM Variations with respect to Sample Size.

maximum possible power of the signal and the power of corrupting noise in it.

- 3) **Structural Similarity Index Measure (SSIM):** SSIM refers to the structural similarity between the two images. In the case of compression, SSIM works as a metric that indicates the change in the compressed image.
- 4) **Space saving (SS):** Space saving shows the amount of space that can be saved using the compressed image instead of uncompressed image. IT can be derived from the equation, $SS = 1 - \frac{\text{Compressed size}}{\text{Uncompressed size}}$
- 5) **Compression Ratio:** It is defined as the ratio between the size uncompressed image to size of compressed image.

C. Evaluation of Learning Process

The evaluation of the learning process is achieved by feeding the system with a set of training samples in batch and measuring the PSNR and SSIM. The batch size is increased step by step for observing PSNR and SSIM. Another experiment conducted by applying the natural images as the test set for the network that is trained using MRI dataset. Figure 4 shows the PSNR variations of a Cross-sectional brain MRI image trained using different sample size batches.

From Figure 4, we observe that PSNR value is proportional to the training batch size. In Figure 4, part a, b and c shows the decoded image for the Cross-sectional brain MRI image.

Figure 3 shows the SSIM and PSNR variations with respect to the training batch size of Sagittal, Coronal and Cross-

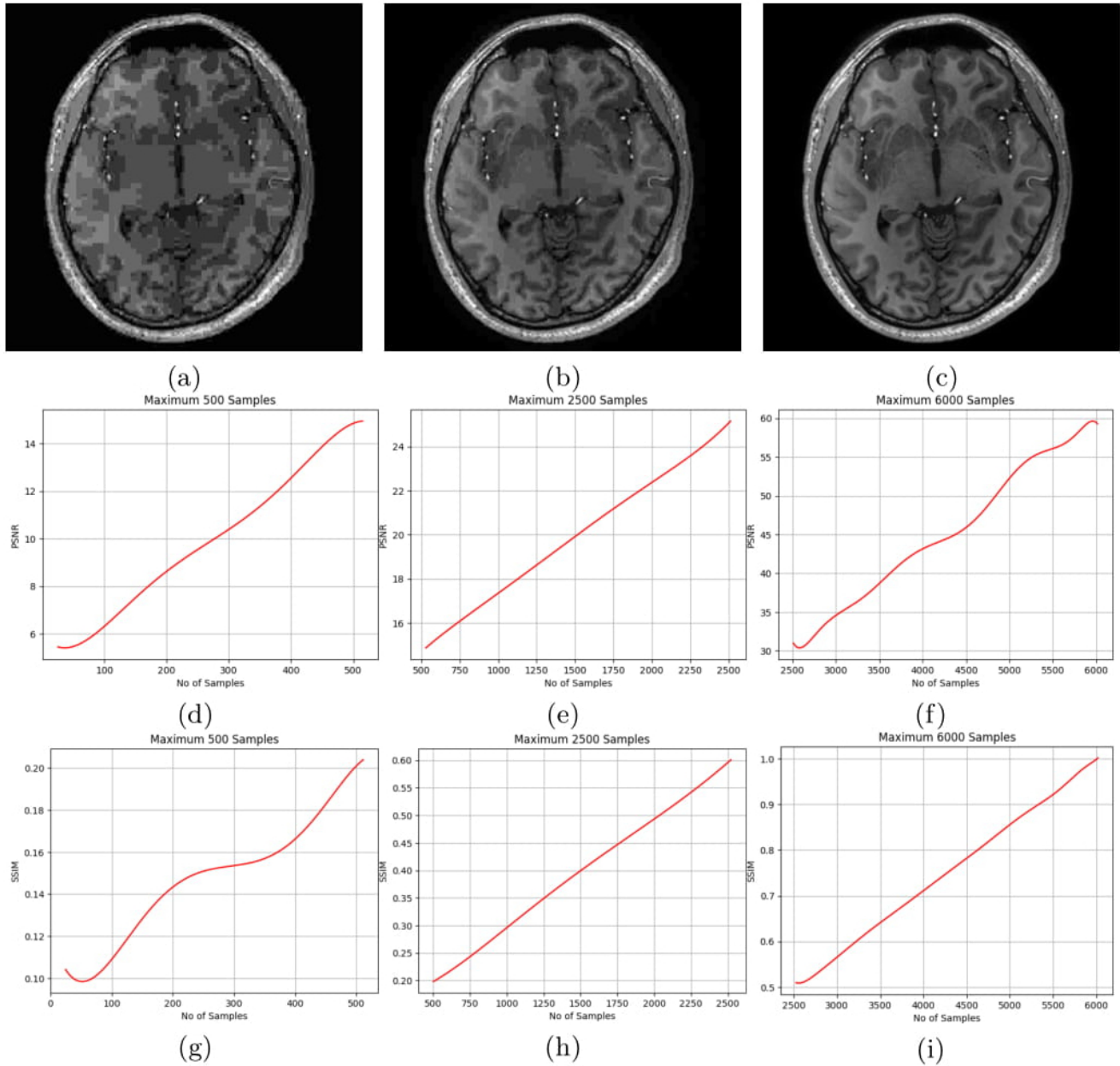


Fig. 4. a)-(c) - Decompressed Image After Training with Sample Size 500, 2500 and 6000 respectively. (d)-(f) - Variations in PSNR with respect to the Sample Set Size 500,2500 and 6000 respectively.(g)-(i) - Variations in SSIM with respect to the Sample Set Size 500,2500 and 6000 respectively.

sectional brain MRI images. It shows that transformation prediction significantly improves by increasing the training batch size. In Figure 5, a cross evaluation of the system is conducted by applying different sets of test images. It shows that system suffers from the over fit towards a trained data set. And the prediction performance of the system completely depends on the category of images used in the training.

So the evaluation process summarizes that the system is able to achieve better results if the training process and testing process uses the same category of images.

System performance is evaluated against four fractal based

compression methods, three classical compression methods and 3 machines learning based compression methods. Evaluation matrices include the PSNR, SSIM, Space Saving, Compression ratio and execution time.

Figure 6 shows the results obtained after training the proposed system with 9000 samples. Performance of the proposed algorithm is compared with three Non-fractal and widely adopted image compression methods-JPEG, PNG and TIFF.

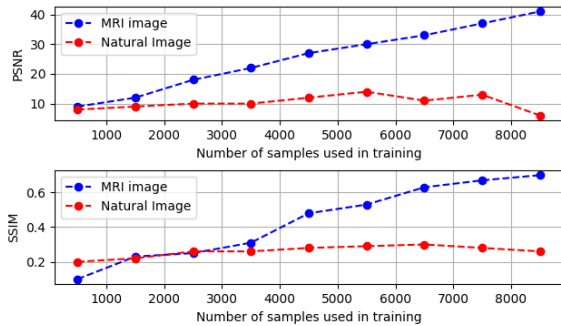


Fig. 5. Cross Evaluation of Proposed System.

D. Evaluation of System Performance

The matrices considered in the comparisons are PSNR and SSIM for quality of the image decompression, and space-saving for storage efficiency of the methods. Part a, b, c in the Figure 6 shows the PSNR, SSIM, and space-saving for different size images. The Proposed method outperforms the other non-fractal methods in all three scenarios. This shows that the suggested method can be a good replacement for the traditional algorithms in an application that requires a large amount of storage space such as Picture Archiving and Communication System(PACS).

Part d, e, f of the Figure 6 shows the comparison between the Machine Learning based Compression algorithms - GMM & Attention [32], iWave++ [31], Non-Local 3D-Context [30]. The proposed method is compared to the existing machine learning method on the basis of PSNR, SSIM and Space Saving. The comparison clearly shows that the proposed method is far superior to the existing techniques.

In Figure 7 the proposed method is compared to the existing fractal methods — Quadtrees [18], No-search [17], Genetic [22]. Part 1 shows that the proposed method outperforms other methods in execution time. So the proposed method solves one of the major disadvantages of the fractal compression, its compression time. Part b shows the proposed method is as good as the existing method in the case of preserving the structural similarity in the image.

VI. CONCLUSION

A time complexity reduction method in fractal image compression has been implemented using the deep reinforcement learning algorithm. While retaining the idea of Iterated function system of domain and range transformations, the search is confined based on a q-learning policy. Consequently, the encoder can compress the image with a better compression ratio and less execution time. This method differs from other recently proposed methods in the prediction of transformations, which doesn't include an exhaustive search. Since the cost to compute the transformations is high, the proposed method uses a neural network-based policy agent to predict the transformations. The empirical analysis shows that the proposed system can be a promising method in the area of medical image compression.

The proposed method identifies DRL as the key technology to reduce the encoding time in FIC. Even if the statistical learning strategy such as DRL is not a popular field in the area of image compression, the proposed work validates the significance of such strategies. This opens up a wide variety of possibilities to modify the image compression techniques with the policy grading algorithms. The user can change the policy grading algorithm depending on the type of images and can take advantage of the DRL based technique to have a better compression ratio with a much higher encoding speed. The Medical image archiving system can be revamped by modifying the existing image archiving system with the proposed DRL based FIC scheme to save the storage space, and hence the cost.

Our work opens up a new area of integrated Machine learning technique for fractal compression. The proposed method can be extended to video and audio compression to achieve a better compression ratio. In the case of video, each frame can be compressed on the basis of initial frame and frame to frame changes can be taken in count to achieve an efficient compression. The method can be used in the case of audio compression, because the number of similar patterns are very high and can be utilized to obtain the better transformation pairs.

REFERENCES

- [1] B. B. Mandelbrot and B. B. Mandelbrot, *The fractal geometry of nature*. WH freeman New York, 1982, vol. 1.
- [2] G. K. Wallace, "The JPEG still picture compression standard," *IEEE transactions on consumer electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.
- [3] P. Aguilera, "Comparison of different image compression formats," *Wisconsin College of Engineering, ECE*, vol. 533, 2006.
- [4] K. Arulkumar, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.
- [5] B. Varghese and S. Krishnakumar, "A novel fast fractal image compression based on reinforcement learning," *International Journal of Computational Vision and Robotics*, vol. 9, no. 6, pp. 559–568, 2019.
- [6] Y. Fisher, E. W. Jacobs, and R. D. Boss, "Fractal image compression using iterated transforms," pp. 35–61, 1992.
- [7] S. K. Mitra, C. A. Murthy, M. K. Kundu, and B. B. Bhattacharya, "Fractal image compression using iterated function system with probabilities," in *Proceedings International Conference on Information Technology: Coding and Computing*, 2001, pp. 191–195.
- [8] S. Liu, W. Bai, N. Zeng, and S. Wang, "A fast fractal based compression for MRI images," *IEEE Access*, vol. 7, pp. 62 412–62 420, 2019.
- [9] M. F. Barnsley, *Fractals everywhere*. Academic press, 2014.
- [10] A. E. Jacquin, "A fractal theory of iterated Markov operators with applications to digital image coding." 1990.
- [11] A. Jacquin, "Image coding based on a fractal theory of iterated contractive Markov operators, Part I: Theoretical Foundation," 1989.
- [12] Y. Fisher, *Fractal image compression: theory and application*. Springer Science & Business Media, 2012.
- [13] B. Hürtgen and C. Stiller, "Fast hierarchical codebook search for fractal coding of still images," in *Video Communications and PACS for Medical Applications*, vol. 1977, 1993, pp. 397–408.
- [14] D. Saupe, "Accelerating fractal image compression by multi-dimensional nearest neighbor search," in *Proceedings DCC'95 Data Compression Conference*, 1995, pp. 222–231.
- [15] Y. Tan and X. Zhou, "A novel speed-up algorithm of fractal image compression," in *International Workshop on Advanced Parallel Processing Technologies*, 2003, pp. 582–589.

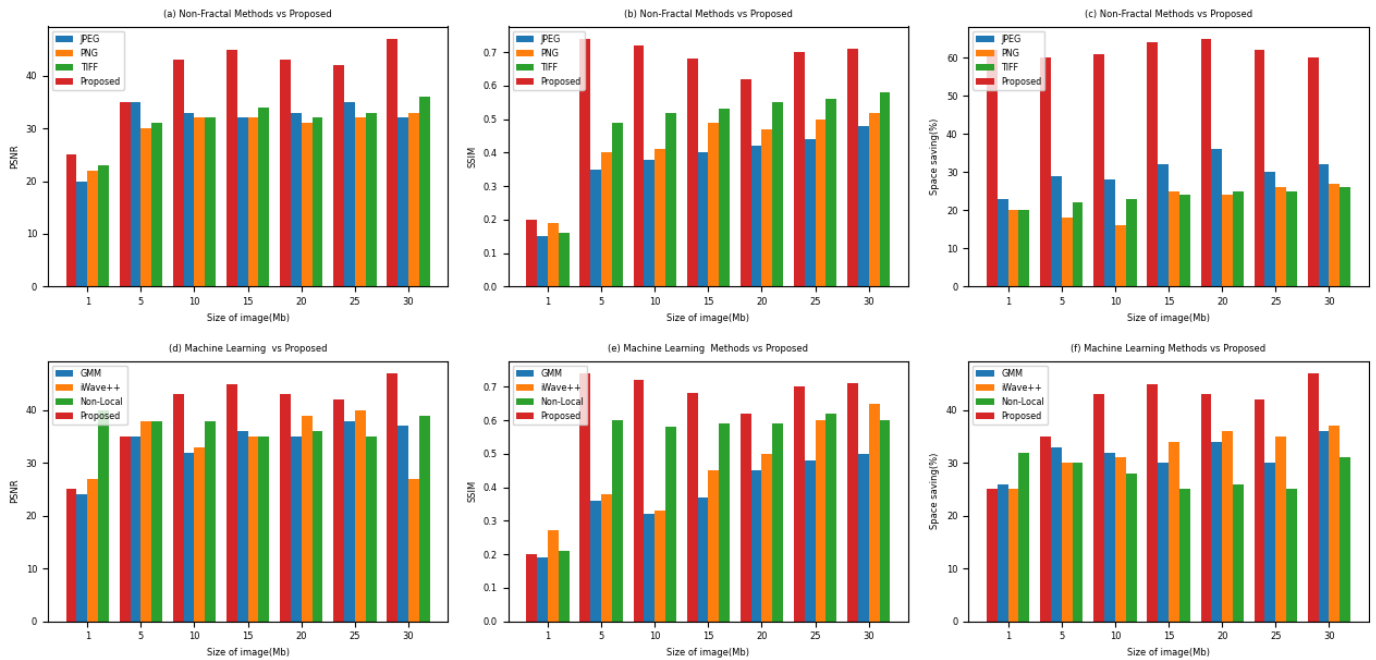


Fig. 6. (a)-(c) - Comparison of Proposed Method and Non-Fractal /Classical Compression Methods. (d)-(f) Comparison of Proposed Method and Machine Learning based Compression Methods.

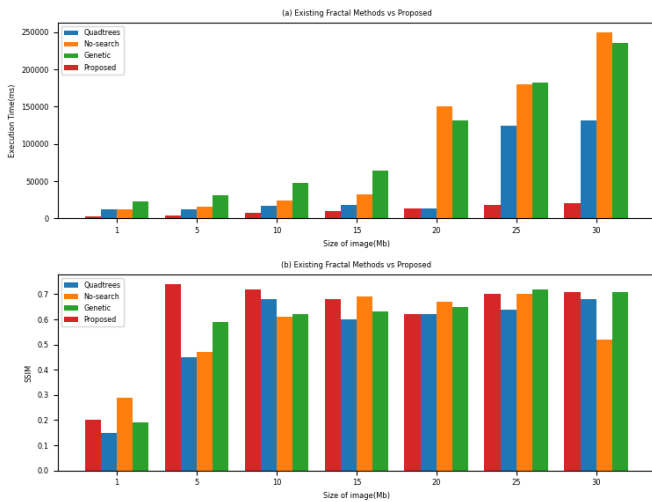


Fig. 7. Comparison of Proposed Method to the Existing Fractal Methods.

[16] X.-Y. Wang, Y.-X. Wang, and J.-J. Yun, "An improved no-search fractal image coding method based on a fitting plane," *Image and Vision Computing*, vol. 28, no. 8, pp. 1303–1308, 2010.

[17] S. Furoo and O. Hasegawa, "A fast no search fractal image coding method," *Signal Processing: Image Communication*, vol. 19, no. 5, pp. 393–404, 2004.

[18] Y. Fisher, "Fractal image compression with quadrees," pp. 55–77, 1995.

[19] V. Chaurasia and S. Sharma, "Similarity based kickoff method for fractal image compression," 2016.

[20] D. E. Golberg, "Genetic algorithms in search, optimization, and machine learning," *Addion wesley*, vol. 1989, no. 102, p. 36, 1989.

[21] J. Kennedy, "Encyclopedia of machine learning," *Particle Swarm Optimization (ed)*, pp. 760–766, 2011.

[22] M.-S. Wu and Y.-L. Lin, "Genetic algorithm with a hybrid select mechanism for fractal image compression," *Digital Signal Processing*, vol. 20, no. 4, pp. 1150–1161, 2010.

[23] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium (PCS)*, 2016, pp. 1–5.

[24] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," *arXiv preprint arXiv:1511.06085*, 2015.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] K. M. A. Okanohara.

[27] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *International Conference on Machine Learning*, 2017, pp. 2922–2930.

[28] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," *arXiv preprint arXiv:1805.11057*, 2018.

[29] M. Sewak, "Policy-Based Reinforcement Learning Approaches," pp. 127–140, 2019.

[30] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "Neural image compression via non-local attention optimization and improved context modeling," *arXiv preprint arXiv:1910.06244*, 2019.

[31] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-End Optimized Versatile Image Compression With Wavelet-Like Transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[32] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7939–7948.

Comprehensive Analysis of Two Malicious Arabic-Language Twitter Campaigns

Reem Alharthi¹, Areej Alhothali², Kawthar Moria³

Department of Computer Science, King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

Abstract—Fake malicious accounts are one of the primary causes of the deterioration of social network content quality. Numerous such accounts are generated by attackers to achieve multiple nefarious goals, including phishing, spamming, spoofing, and promotion. These practices pose significant challenges regarding the availability of credible data that reflect real-world social media interactions. This has led to the development of various methods and approaches to combat spammers on social media networks. Previous studies, however, have almost exclusively focused on studying and identifying English-language spam profiles, whereas the problem of malicious Arabic-language accounts remains under-addressed in the literature. In this paper, therefore, we conduct a comprehensive investigation of malicious Arabic-language campaigns on Twitter. The study involves analyzing the accounts of these campaigns from several perspectives, including their number, content, social interaction graphs, lifespans, and day-to-day activities. In addition to exposing their spamming tactics, we find that these spam accounts are more successful in avoiding Twitter suspensions that has been previously reported in the literature.

Keyword—Social network security; social spammers; arab twitter users; malicious campaigns on twitter; data mining

I. INTRODUCTION

Social media networks have profoundly affected life today and have become massive platforms for communication and information exchange. A remarkable example of this influence is the role of social media in the Arab Spring, which was extensively reported in the literature [1], [2], [3]. People have embraced such web platforms as independent media not subject to political parties or organizations. Today, in a highly volatile political setting, social media networks continue to play the same critical role in the Arab world. Researchers also utilize this social media data to extract information for various objectives such as opinion mining for the Arabic language [4], event detection [5], [6], and rumor detection [7].

The number of Arabic-language users on social media networks, including ill-intentioned individuals, increased tremendously after the Arab spring [8]. Some users misuse these websites to share inappropriate, deceptive, and offensive material, for intrusive advertising, public opinion manipulation, and to spread malicious malware, for example. They usually manage an enormous number of accounts known as campaigns and employ different spamming strategies to achieve their goals. Besides restricting freedom of expression, the quality of social media content as an informative tool used for various purposes is inevitably diminished by these actions. There exists an extensive literature on this topic [9], indicating that

the problem of malicious users or campaigns is not recent; however, for Arab users, few studies have been presented to detect such activities at the account and individual tweet levels. Although some attempts have been made to analyze Arabic-language social media spammers [10], [11], this area is still insufficiently explored, as only traditional spammers have been investigated. Besides this and to our knowledge, there is no previous work investigating Arab spammers at the campaign level or investigating various aspects such as their spamming strategies. Therefore, an extensive analysis study has particular importance given the recent aggressive activity of malicious campaigns in the Arabic-language Twittersphere, where every day, malicious users flood trending topics with a tremendous amount of spam and low-quality content (more details in Section IV-B3).

This paper presents an empirical analysis of two malicious Arabic-language campaigns on Twitter. Compared to other groups in our dataset, they have the highest numbers of fake accounts and the longest periods of such harmful activities. The overall goal of this analysis is to extensively investigate the content and behavior of the two groups, including their numbers of accounts, tactics, lifespans, and methods used to control their large numbers of profiles. Accordingly, six hundred profiles of the two campaigns have been analyzed and split into two generations: one generation's activities were recorded for Apr-Dec 2018 and the second one's for Aug-Sep 2019. Through this analysis, many results are revealed regarding malicious Arabic-language groups, including how they coordinate accounts in clusters, how they use the clusters to target trending topics, their group interaction characteristics, the content characteristics described by the tweet functions, and the self-similarity ratios of the profiles. This study also addresses the expected lifetimes of such accounts as well as how they manage to run a large number of accounts (i.e., manually or using software).

Having greater insight into the activity of malicious campaigns will provide useful information about the quality and credibility of social network data. Such information is essential not only for the spam detection field but also for other areas that rely on social network data. Thereby, this study attempts to provide an in-depth insight into the current situation of Arabic-language trending topics and malicious campaigns, and more specifically, their strategies for abusing trending topics to maximize the distribution of their content as well as identifying the primary characteristics of such accounts and comparing them with the features of spammers as reported in the literature. In addition, through a two-month experiment,

this study tries to analyze the lifespans and daily activities of the campaign accounts, as well as the Twitter system's ability to detect Arabic-language social media spam accounts.

The rest of this paper is structured as follows: Section II describes previous related works, and Section III discusses the dataset collection process. Section IV is divided into three major sections: the first introduces the number of campaigns' accounts (Section IV-A), the second provides an in-depth analysis of the main groups' characteristics, including the groups' interaction graphs, spamming strategies, and content attributes (Section IV-B), and (Section IV-C) discusses the practice of managing spam accounts. Section V summarizes the findings of this in-depth study. Section VI provides a summary of this paper, including the conclusion and suggestions for future work.

Finally, we note that a part of this paper appeared previously as a conference publication [12]. This part was included in the Data Analysis section in the conference paper, in which we briefly presented the clusters' organization and the automated behavior of the campaigns' accounts. Our main contributions for the journal version include an expanded detailed analysis that covers several aspects of these groups.

II. RELATED WORKS

A. Malicious Campaign Studies

Detecting individual malicious accounts on the scale of an entire social network is a costly and time-wasting approach. A study [13] highlighted the importance of addressing malicious accounts at the group level, which is often a more feasible and effective solution. The group-level detection approach identifies campaign accounts according to their common materials or objectives, which in turn raises the bar for the attackers to evade detection. Creating unique content or running every single account separately to hide the similarity among the group would greatly increase the costs and time to administer these accounts; therefore, researchers have suggested several systems and strategies expose various types of malicious accounts at the campaign level. For instance, a study [14] proposed examining the social graph between users and pages to reveal Fake-Likes campaigns on Facebook, and several studies have used the synchronized behavior and timing of social spammers' fraud activities, fake Twitter followers, and malicious retweeter groups to expose their accounts on Twitter [15], [16], [17], [18], [19]. Besides, a variety of analysis studies have been carried out to understand the various aspects of social spammers. Yang et al. [20] examined spammers' social graphs to identify the relationships and supporters of these accounts and showed that they are socially connected in communities and, in a number of cases, are supported by legitimate accounts. A study [21] investigated spammers' strategies to enhance their influence scores by following real users as well as each other on Twitter. Lastly, Gupta et al. [22] conducted a large-scale analysis study of spam campaigns on multiple social platforms that used telephone numbers to lure victims.

B. Detecting Spam in Arabic Social Networks Content

Most of the literature has focused mainly on English-language spammers and has made fewer attempts for non-

English language users, even though spammers' strategies probably vary from one region to another given the fact that they evolve and find a new spamming method over time [23]. This section provides a brief overview of the body of related work, with a focus on detecting Arabic-language spammers in social networks.

One of the earliest empirical analyses of Arabic-language spammers presented by Al-Khalifa et al. [10], in which the authors examined the content and social graphs of these accounts, showed that they are still naive. A comprehensive investigation conducted by [11] studied the characteristics of long-surviving but eventually suspended Arabic Twitter accounts, and in this study, the authors compared these accounts with short-lived suspended accounts in terms of their content, activity, and linguistic attributes. Accordingly, they found that the short-lived group had a high self-similarity ratio compared to the long-lived group. Regarding the degree of activities such as gaining more followers or friends and posting tweets, the long-lived group was more active, and meanwhile, they avoided excessive behavior such as posting large numbers of tweets.

Research by [24] reported that spam tweets constitute about three-quarters of Saudi Arabia's trending tweets. The study also assessed the efficiency of well-known features that are designed based on English spam profiles in detecting Arabic spam accounts. First, they selected a range of features that combine profile and content characteristics to reflect the reputation and replication characteristics of an account, and then they compared the performance of the selected features and the model and features proposed by [25] that was also tested on the Arabic dataset. The results indicated that the selected features performed better than the model proposed by [25] in detecting Arab spam profiles. To classify spam at the tweet level, the authors in [26] designed a classification scheme that used content-based features such as the number of URLs, hashtags, phone numbers, and spam words present in a tweet.

Considering that automation technology can be used for malicious purposes such as spamming, an attempt by [27] was made to expose automated Arabic tweets. The proposed system tested different factors to identify a tweet as an automatic or manually generated tweet. In addition to the degree of formality of the tweet, several structure-based features such as the length of the tweet are employed in the classification decision. Nonetheless, automated tweets need not necessarily be malicious tweets, as there were no specific rules in the article to differentiate between malicious and benign automated tweets.

III. DATASETS COLLECTION

To collect a sufficient set of data, a Python crawler that uses Twitter API functions is developed to gather the required information. The crawler was first used to randomly collect tweets and users from Saudi Arabian trending topics, from which we excluded non-Arabic profiles. Using the crawler over 4,000 different identifiers and 160,000 tweets are assembled. We then manually annotated 1,000 legitimate accounts out of the random set of users. To investigate malicious campaigns, a collection of accounts exhibiting spam-like behaviors such as sharing duplicate tweets and

URLs were also gathered from trending topics in Saudi Arabia. Following that, we looked at the content and practices of these accounts, classifying individuals with similar or duplicate materials as a group, which is a sample strategy used by several previous studies [22], [28], [29]. The dataset eventually located two campaigns that stood out significantly from the rest of the collected groups. The first campaign consisted of 200 spamming accounts that shared a duplicate URL to an external website. The second group also had 200 accounts, and their work focused on the promotion of unfamiliar, unlicensed medicinal brands. Aside from having the longest periods of such harmful activities, two campaigns were discovered to have the highest number of fake accounts when compared to other groups. We, therefore, chose to focus our attention on these two campaigns (spammers and promoters). In addition, 200 accounts from second-generation (spammer and promoter) campaigns are added to the dataset in order to investigate them in depth and track their behavior over time.

IV. IN-DEPTH ANALYSIS

A. The Number of Campaigns' Accounts

Our preliminary analysis of the number of accounts corresponding to the spammers' and promoters' campaigns shows that at any given time, they operate numerous accounts to spread their content, and that suspended or deleted campaigns' accounts will constantly be replaced by a new generation. As shown in Table I, the spammers' accounts (S_{G_1}), which had activity from Oct to Dec 2018, were all eventually suspended by Twitter, and the attackers replaced them with (S_{G_2}). The same thing goes for the promoters' group (P_{G_1}) and their second-generation (P_{G_2}). The accounts shown in Table I comprise the total number of accounts analyzed in the course of this study and do not reflect the actual number of campaign accounts.

For both campaigns, as Table I shows, the age of the accounts is above 1, 500 days (4 years) on average. By contrast, several studies for non-Arabic spammers found that these accounts tend to have a young age, of less than 200 days [25], [30], [31]. The long lifetime of the spammers' accounts implies that these accounts are compromised accounts, i.e., accounts that have been stolen from legitimate owners and which exhibit dramatic changes in behavior and content patterns [32], [33]. To confirm that, the account's tweets and profile are manually inspected to see if there was any evident behavioral change such as excessive posting rate, sharing spam URLs, and sharing duplicate content. The following points summarize the results of the experiment:

- According to the findings of this study, the majority of campaign accounts shared duplicated tweets in their most recent activities, whereas the first tweets were genuine tweets that did not include spam or duplicated links. In addition to the duplicated content, the tweet source, which is the utility used to post the tweet, was the most common sign of behavioral changes that we observed in our dataset.

- Through additional content analysis, two classes of account are identified; the first class includes accounts in which the old tweets (genuine tweets that are not duplicated, nor contain a spam link) are still there, and represents about 36.52% of the accounts in our dataset, and the second class involves accounts in which the old tweets were deleted (approximately 63.74% of the accounts). For the second class, we found several accounts that had explicitly used some service (such as TweetDelete) to automatically delete all the old tweets¹.
- By analyzing the activity timelines of the campaign accounts, significant time gaps in the histories of these accounts can be observed which range between six months and four years; in the case of profiles in the first class particularly, the time between the last genuine tweet and the first spam tweet (Figure 1(a)), and in the case of accounts in the second class, the time between the account's creation date and the first tweet date, as shown in Figure 1(b). As Figure 1 shows, the creation date and the last genuine tweet most frequently occurred before 2018, and all the spam tweets were sent by the end of 2018 or in 2019.
- Figure 2 shows an example of a compromised account used in spamming activity. There is a significant difference in the behavioral patterns of the account's tweets before 2015 and the tweets in 2018, which involve duplicated tweets and using a different tweet source.
- The time gaps and the accounts' behavioral changes provide clear evidence that most of the accounts used by the two campaigns are compromised accounts.

B. The Main Characteristics of Malicious Groups

The results for the main groups' characteristics are presented in the subsections below, where various analytical methods are used to investigate the groups' interaction graphs, spamming strategies, and content attributes.

1) *Groups' Interaction Graphs*: In contrast to traditional spammers who aggressively and randomly share unsolicited content, both the promoters' and spammers' groups have shown high organizational levels. The two groups manage their large numbers of accounts in small-scale clusters of connected accounts, with about 4 to 18 profiles in each cluster. Individual clusters work autonomously toward a specific goal, e.g., each group in a promoters' campaign promotes a particular medicine with the same brand name. Similarly, in a spammers' campaign, each cluster takes advantage of specific trending stories and encourages people to follow a hyperlink (a spam URL) to learn more details about the trending topic. Despite that, not all the groups are entirely independent in terms of their content, as we found that some of the groups share similar content.

¹ <https://tweetdelete.net/>

TABLE I. GROUP DESCRIPTIONS AND METADATA STATISTICS

	Activity Type	Activity Duration	Number of Accounts	Account Age	Followers	Friends
(S_G ₁)	Spreading spam URLs	Oct-Dec 2018	200	1,604 days	116.32	236.85
(S_G ₂)	Spreading spam URLs	Aug-Sep 2019	100	1,511 days	178.35	142.43
(P_G ₁)	Promoting unlicensed medicines	Apr-Oct-Dec 2018	200	1,894 days	53.29	134.95
(P_G ₂)	Promoting unlicensed medicines	Aug-Sep 2019	100	1,510 days	138.98	416.32

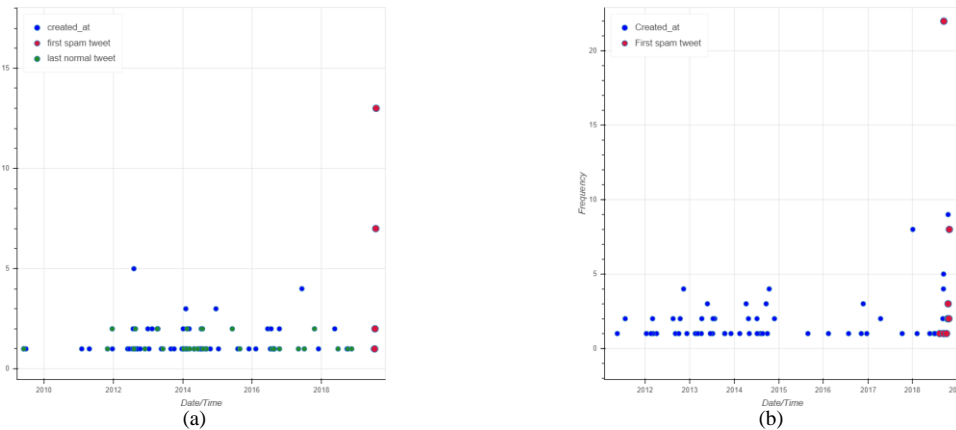


Fig. 1. Class 1. Accounts for which Old Tweets were not Deleted; (b) Class 2. Accounts for which Old Tweets were Deleted.

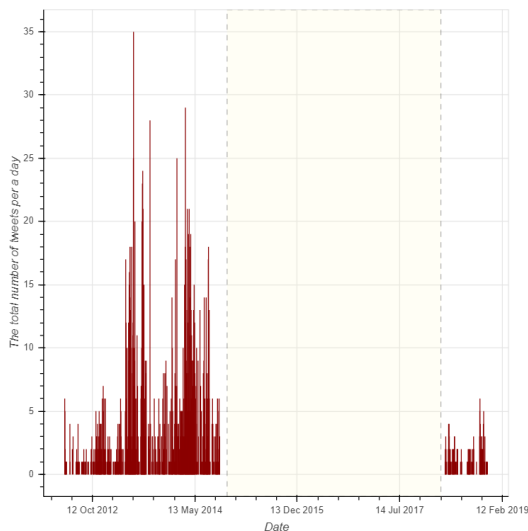


Fig. 2. An Example of a Spam Account Activity over Time; Last Genuine Tweet was on Oct 1, 2014, and First Spam Tweet Occurred on Mar 26, 2018. Twitter for iPhone is the Source of the Last Genuine Tweet, and Twitter Web Client is the Source of the Spam Tweets.

A distinct spamming strategy is identified by studying the campaign clusters and their account interactions. This strategy is mainly developed to invade and flood hashtags and trending topics with unsolicited tweets. The accounts are organized and assigned to a specific social role inside a cluster, as follows. The clusters involve one or more central accounts whose role or task is posting original (spam or promoting) tweets. Then, a set of accounts regularly retweet, replay, and generally interact with whatever the central accounts share. Attackers with this strategy are guaranteed to reach a large group of users, especially in trending topics. This tactic will also promote spam to the “best tweets” tab by manipulating

Twitter’s tweet-rating tools, which assess tweets according to overall engagement, for instance, the number of retweets, likes, or replies. It is worth mentioning that the “best tweets” tab is the default tab for trending topics and the place where people often look for the most influential tweets.

In a general sense, the campaign accounts work or interact together within clusters. To visualize their organized behavior and interactions, a campaign interaction graph is constructed, which is defined as $G=(V,E)$, where V is the graph’s vertices/accounts and E is the edges that connect two vertices if there is an interaction between them [23]. In the interaction graph, three types of interaction between accounts is defined: retweets, replies, and mentions. We utilized the Networkx package² to build an undirected graph that shows the accounts’ interactions from the topological point of view. The observations about the interaction graph are discussed in the following points:

- Because the clusters are often independent of each other and follow the same strategy, we have chosen to visualize the interaction graph at a cluster level rather than visualize all the campaign accounts’ interactions. In addition, the interaction graph for three clusters is constructed to demonstrate the differences between the genuine and spammer classes: spammers, promoters, and genuine, with equal numbers of tweets and accounts. The clusters’ interaction graphs were built according to their accounts’ most recent 20 tweets ((Figure 3 (a), Figure 4 (a), and Figure 5 (a)) or their overall tweets ((Figure 3 (b), Figure 4 (b), and Figure 5 (b)) (see Figure 3 (a ,b), Figure 4 (a ,b), and Figure 5 (a ,b)).

² <https://networkx.github.io/>

- The groups organized behavior is very clearly shown in Figure 3 (a), and Figure 4 (a), in which ac-accounts intensely interact with the central accounts in the clusters (the red node is the main central account). For the spammers' cluster, 26 nodes/accounts and 144 edges/interactions for nine accounts are found in their last 20 tweets, and similarly, there are 21 nodes/accounts and 104 edges/interactions in the promoters' graph. In contrast, in the genuine graph Figure 5 (a), there are 49 nodes/accounts and 49 edges/interactions, which indicates a more genuine or organic behavior.
- To assess graph connectivity, Table II presents the average clustering coefficients of the spammer and promoter groups' graphs. A clustering coefficient (CC) is a measure that indicates if the graph nodes are part of a highly connected graph [34]. Despite the large number of edges/interactions in the spammer and promoter groups', the average CCs of the accounts is zero, which indicates that their graph topology is a star. In other words, all the accounts' interactions are directed to the central accounts, and the central accounts do not interact with each other.
- As stated in Table II, the graph diameter of the spammer and promoter groups is equal to two, which indicates that there is a node that connects with every other node in the graph. Clearly, the central accounts are the nodes that connect all the other accounts, which is also reflected by the maximum degree of centrality and maximum closeness centrality properties of the groups' graphs.

Remarkably, we found a few isolated spam and promoter accounts that operated as a single account and posted tweets without further interaction with other accounts. Also, several promoters' accounts that had a strategy different from what has been discussed in this section are found. They abuse the "mention" function to reach a particular audience or user in the trending topics rather than many users. More specifically, they mention or reply to popular accounts or top tweets in a specific topic with their promotion tweets, which is a well-known strategy for spammers [35].

2) *Groups' content attributes:* Several previous studies: have highlighted the importance of content-based features [31], [36], [37] in identifying social media spammers. Generally, the content or language model of the spammers is

significantly different from genuine accounts' content as a result of their distinct ill-intentioned use of social networks. In this way, they attempt to maximize their content distribution by intensively posting duplicate texts and URLs or aggressively exploiting the network's services, e.g., hashtags, mentions, URLs, and photos. Therefore, in this section investigates the content characteristics of Arabic spam campaigns in two aspects: using the tweet functions and the self-similarity and word frequency of the spammers' language.

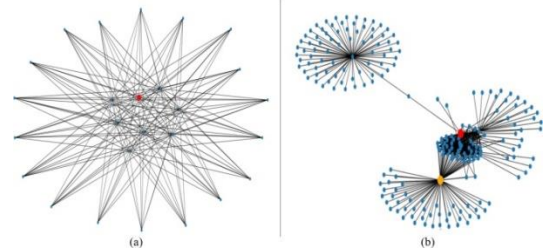


Fig. 3. (a) The Interaction Graph of 9 Spam Accounts according to their Most Recent 20 Tweets. (b) The Interaction Graph For the Same Spam Group according to all their Tweets.

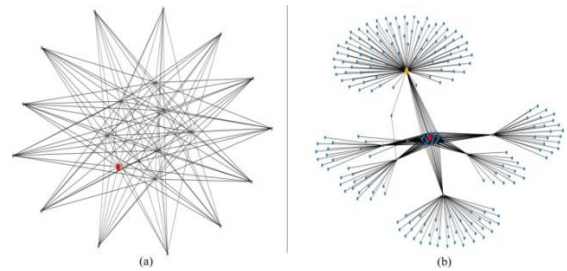


Fig. 4. (a) The Interaction Graph of 9 Promoters' Accounts according to their Most Recent 20 Tweets. (b) The Interaction Graph for the Same Promoters' Group according to all their Tweets.

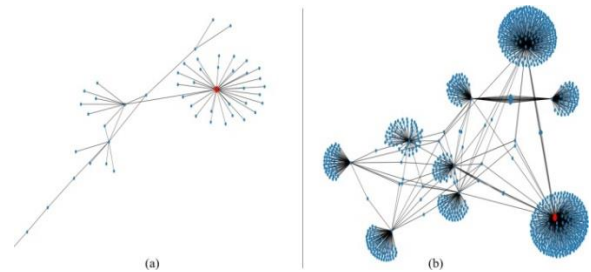


Fig. 5. (a) The Interaction Graph of 9 Genuine Accounts according to their Most Recent 20 Tweets. (b) The Interaction Graph for the Same Genuine Group according to all their Tweets.

TABLE II. GROUPS' GRAPH PROPERTIES

	Spammers		Promoters		Genuine	
	(a)	(b)	(a)	(b)	(a)	(b)
Clustering Coefficient	0	0	0	0	0.0292	0.00889
Average Degree	11.1	4.9	9.9	3.1	2	2.1
Graph Diameter	2	6	2	4	7	6
Maximum Degree Centrality	0.72	0.42	0.65	0.5	0.61	0.31
Maximum Closeness Centrality	0.78	0.41	0.74	0.51	0.56	0.46
Maximum Betweenness Centrality	0.06	0.59	0.05	0.66	0.84	0.52

Various statistics from the accounts' tweets for the content attributes are first collected, such as the numbers of URLs, photos, and hashtags. Figure 6 (a, b, c, and d) plotted a cumulative distribution function (CDF) for the content attributes to compare between the spammer and promoter groups and the genuine accounts in our dataset. The points listed below discuss the findings:

- Spammer groups exhibit aggressive behavior in using most of the tweet functions, in which they post many links, hashtags, and photos per tweet. Additionally, their content attributes vary considerably from the genuine accounts and promoters' accounts, as shown in Figure 6 (a, c, and d).
- Promoters' accounts show similar patterns across multiple attributes to the genuine accounts as opposed to spammers, as shown in Figure 6 (a, b, and c).
- As shown in Figure 6 (d), the unique URL ratio exhibits the highest divergence between the three classes in our dataset.

To estimate the semantic similarity of pairwise tweets for the campaign accounts, we take the average of the word embeddings of all words in the pair tweets and then compute the distance between the resultant sentences' vectors by using cosine similarity. The pre-trained word2vec model [38] is used to obtain the embedding vectors of the words. As shown in Figure 7 (a), the self-similarity [11] of 40% of both the spammers' and promoters' accounts is greater than 0.7, while less than 1% of the genuine accounts reach the same percent of similarity. Additionally, the new generations of the campaign

accounts follow almost the same distributions as the old suspended accounts, as shown in Figure 7 (b). The high self-similarity ratio suggests that these accounts are designed to deliver or distribute one message, which conforms with our findings in the previous section. More precisely, these accounts concentrate on promoting, for example, a particular service or medicine in the case of promoters' accounts or taking advantage of a specific controversial story in the case of spammers' accounts.

- In addition to their high self-similarity, the spammers' and promoters' accounts often use the same sets of words to deliver their messages. Figure 8 shows the average of newly introduced words in the accounts' tweets through-out 100 tweets. As Figure 8 shows, genuine accounts use 6 to 8 new words in each new tweet, while both the spammers and promoters introduce two new words over their new tweets, which are more likely to be either keywords or hashtags in trending topics.
- A further interesting observation is that spammers' campaign tweets mostly relate to trending topics or hash-tags. For example, if there is a trending story or viral news, these accounts usually claim that their spam URLs provide more information about the story. This stands in contrast to many previous studies that have described the spam tweets as tweets that are irrelevant to the topics [9],[23], [35]. In general terms, both campaigns post content that is easily detectable and varies from the material of real users.

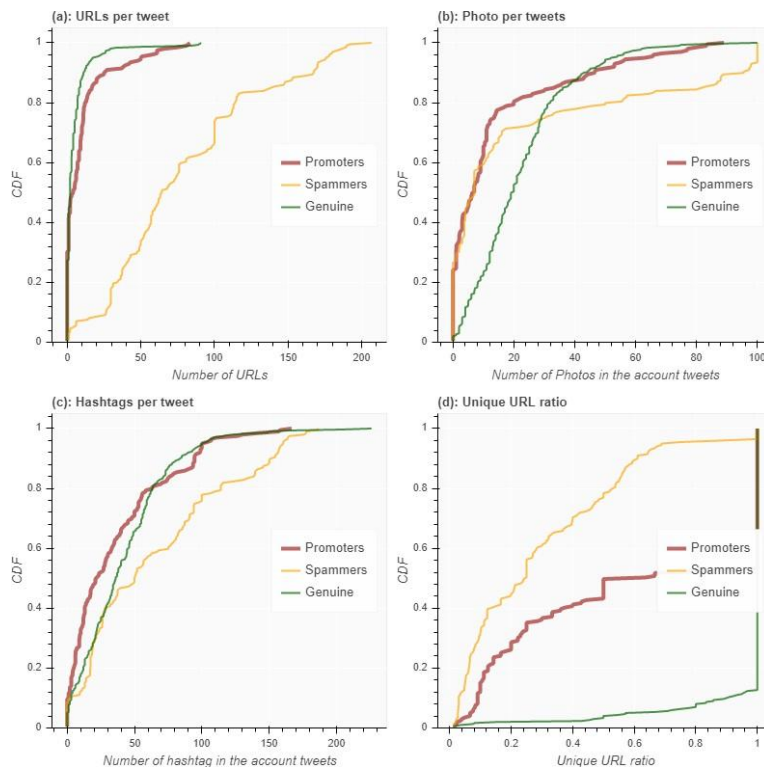


Fig. 6. (a). Number of URLs Contained in the Accounts' 100 Tweets. (b). Number of Photos in the Tweets. (c). Number of hashtags Contained in the Accounts' Last 100 Tweets. (d). Number of Unique URLs vs. Number of Shared URLs.

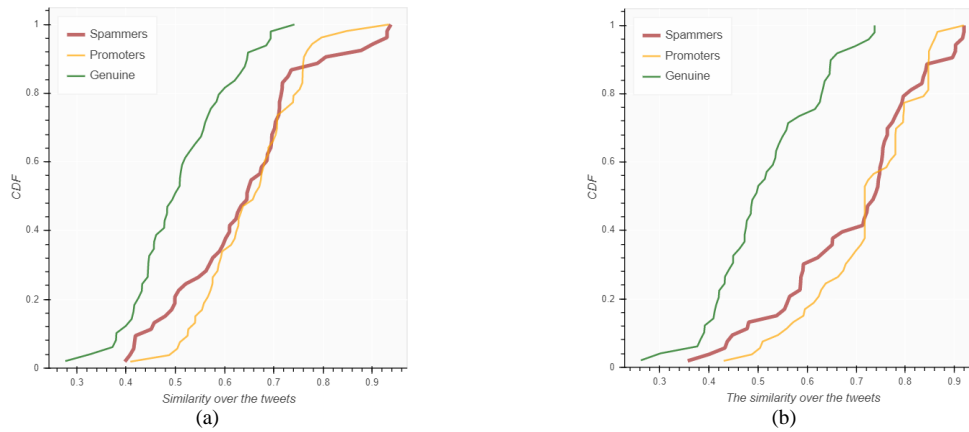


Fig. 7. (a). Comparison between the Three Classes' Accounts (Spammers, Promoters, and Genuine) in Terms of their Self-Similarity Ratios. (b) Comparison between Old Suspended Generation of Campaigns and New Active Accounts.

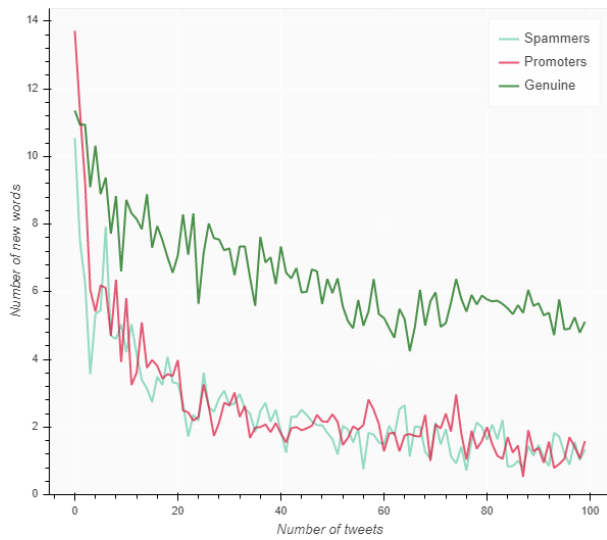


Fig. 8. Average Number of New Words over the Accounts' Last100 Tweets.

3) The campaigns' accounts lifespan and daily activities:

As previously mentioned in Section IV-A, the two campaigns constantly replace the suspended accounts with new ones. As a result, they have the most extended durations of spamming activities and the most significant numbers of fake accounts (see Section IV-A). This section addresses the accounts' lifespans in order to answer the following questions:

- (1) How do attackers leverage these accounts to obtain the maximum possible output?
- (2) what are the average lifespans of these accounts?
- (3) how can these accounts avoid the Twitter detection system?
- (4) among the different kinds of spammer and promoter activities, which ones are the fastest to be detected by Twitter? And
- (5) among the spammer and promoter accounts, which ones can survive suspension for more extended periods of spamming activity?

To answer these questions, we conducted a two-month investigative study of the second-generation of the spammers' (S_{G_2}) and (P_{G_2}) promoters' groups (see Table I). During the two months, we followed or recorded the daily activities of these accounts, including their new tweets (i.e., original,

retweet, and mention tweets), numbers of total tweets, numbers of followers and friends, and the current state of the accounts (i.e., active or suspended). Additionally, to examine their social interaction patterns, we tracked new tweets, total received retweets, and favorites.

Over the two months, the 200 accounts of the (S_{G_2}) and (P_{G_2}) produced a total of 243, 037 low-quality tweets, as illustrated in Figure 9. The spammers' group accounted for a large percentage, at nearly 72% of total spam tweets. The fact that this vast number of tweets was generated by a subset of the actual number of campaigns' accounts is particularly alarming. We believe that the two campaigns had a much larger number of fake accounts, from which the 200 accounts are collected by searching a few keywords and trending topics. This massive number of spam tweets reflects the current crisis of Saudi Arabian trending topics, where such accounts flood the trends with disturbing and unsolicited content on a daily basis [39]. Twitter, on the other hand, had succeeded in suspending 55% of the campaigns' accounts over the two months but failed to identify about 45% of the total number, which were still active up to the last day of the experiment.

The average, median, and maximum number of tweets per day for the suspended and active accounts are computed to compare the posting ratios of the two groups. In addition, the posing ratios of the genuine accounts is estimated by computing these statistics for 1, 000 labeled accounts from our previously collected dataset (see Section III). As shown in Table III, the spammers continued to exhibit aggressive behavior, with the highest posting ratio among the three classes. Surprisingly, the active of (S_{G_2}) accounts were more aggressive than the suspended group, with an average of 90.21 tweets and an account share of over 6, 143 spam tweets per day. The promoters group (P_{G_2}) again showed a pattern that closely matched the genuine accounts, with an average of 22.57 tweets per day. Regarding the groups' social graphs, almost 90% of the accounts maintained the same numbers of followers and friends, and the rest of the accounts' numbers actually decreased during their spamming periods. That indicated that these accounts were aimed at targeting a broad audience in the trends instead of using the "follow" function to reach for specific victims.

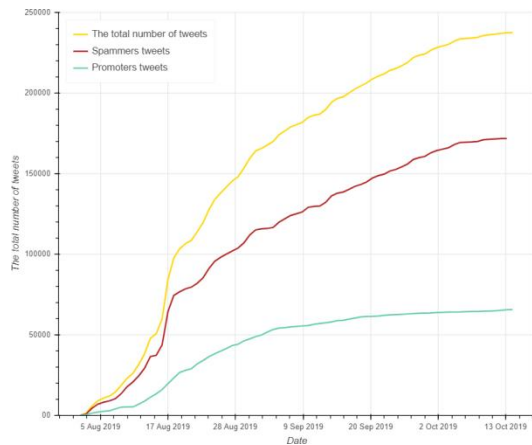


Fig. 9. Total Number of (S_{G2}) and (P_{G2}) Tweets over Two Months.

TABLE III. DAILY AVERAGE OF TWEETS FOR THE GROUPS' ACCOUNTS AND GENUINE ACCOUNTS

		Average	Median	Maximum
(S_{G2})	Suspended accounts	59.22	29.5	707
	Active accounts	90.21	37.5	6143
(P_{G2})	Suspended accounts	31.11	25	229
	Active accounts	22.57	12.75	396
Genuine Accounts	-	23	8	100

We defined the lifespans of these accounts as the duration between the date of the first tweet that violated Twitter rules and the date of the last tweet. We believe that this provides a more precise definition of malicious account activities than the interval between the account creation date and the date of the latest tweet, since the interval time would incorporate all an account's activities, including its genuine early stage in the case of compromised accounts (see Section IV-A). Additionally, this definition gives a more accurate description of how long these accounts can avoid suspension after their first spam tweets.

Table IV provides the average, median, and maximum duration of activities in days for the second-generation of spammer and promoter accounts. In the case of suspended accounts, the duration is the time between the first spam tweet (most of these tweets occurred after Aug 1, 2019) and the date of the last tweet before the account is suspended. For the active accounts, the duration is the time between the first spam tweets and the latest tweet shared by the account, and the accounts were still active until the last day of the experiment. As shown in Table IV, the average lifespan of the (S_{G2}) accounts in the suspended group was less than that of the suspended (P_{G2}) accounts, which is an expected outcome due to their aggressive behavior. Also, Twitter could detect the (S_{G2}) accounts faster than the accounts of (P_{G2}), as shown in Figure 10. Even though the active (S_{G2}) accounts had a longer average lifespan than the (P_{G2}) active accounts, and their total number of accounts was at some point greater than the (P_{G2}) accounts during the experiment (see Figure 10), the promoters' campaign was more successful in leveraging their accounts. In consequence, the spammers'

campaign exhibited very easily detectable behavior, as a result of which 60% of their accounts were suspended either on the first day or in less than five days (see Figure 13 (c)). Furthermore, this study discovered that the long-lived spammers' accounts were "sleepers," which meant that their activities would pause for a short period of time. The promoters' accounts, on the other hand, posted fair numbers of tweets daily; consequently, their activities (including their tweets that were still public in the trends) persisted for a more extended period than those of the spammers. As shown in Figure 13 (b and d), only 18% of the promoters' accounts were suspended within 5 days of their activities.

Also, the sets of the suspended accounts in both campaigns are examined to clarify the relationship between the accounts' lifespans and posting patterns. The primary assumption of this experiment was that accounts with high posting ratios would be more likely to be identified than those with lower posting ratios. Figure 11 shows the average posting ratios per day plotted against the lifespans of the suspended accounts. Many promoters' profiles shared 30 tweets a day on average, and lived almost 30 days, as Figure 11 shows. For spammers' accounts, they were more likely, regardless of their posting ratio, to be identified within less than ten days. That situation, however, does not extend to all campaign accounts, which might be a result of other factors contributing to the suspension. For instance, accounts might be suspended after being reported or flagged for containing disturbing or spamming content by genuine users.

TABLE IV. SPAMMERS' AND PROMOTERS' ACCOUNTS' ACTIVITY DURATION IN DAYS

		Average	Median	Maximum
(S_{G2})	Suspended accounts	16.75	3	190
	Active accounts	42.56	16.5	269
(P_{G2})	Suspended accounts	20.22	23	51
	Active accounts	21.18	6.5	73

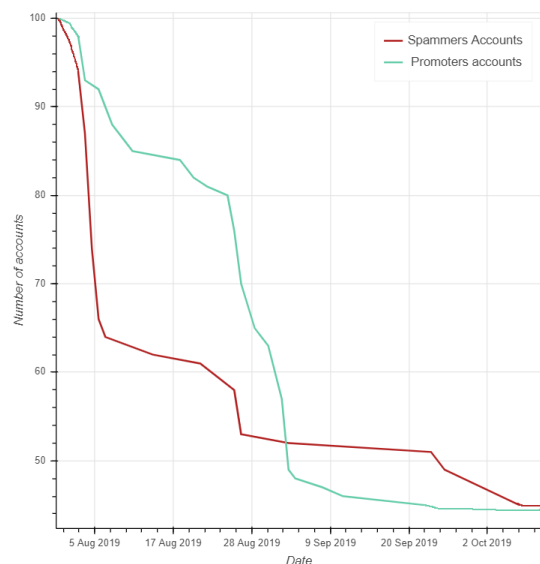


Fig. 10. Illustration of Total Number of (S_{G2}) and (P_{G2}) Accounts over the Two Months.

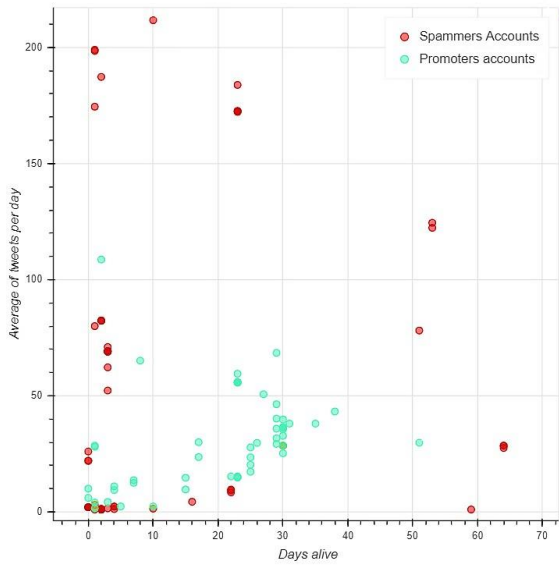


Fig. 11. Duration of Suspended Accounts' Activity and the Daily Average of Tweets.

Similarly, the activities of accounts in the suspended set are examined and compared them to those of the active accounts of the two campaigns. The purpose of this analysis is to understand which account behaviors were the most detectable, e.g., retweets, mentions, and centering accounts (see Section IV-A). However, we could not adequately compare all the behaviors due to the small number of second-generation accounts and the different samples in each group (see Figure 12). Comparing our analysis results with previous studies on Arabic-language spammers [10], [11], we first found that the spam accounts' lifespans in our dataset were much longer than reported in other studies. In [11], for example, it was found that 50% of accounts were detected and suspended after their first spam tweet. In our case, only one account in the promoters' group was detected after three tweets, while the rest of the suspended accounts shared 6 to 1,780 tweets before suspension (see Figure 13 (b)). For the spammer groups, only 2% of total accounts were detected after five or fewer tweets, while the rest shared through their spamming period from 6 to 6,101 tweets (see Figure 13 (a)).

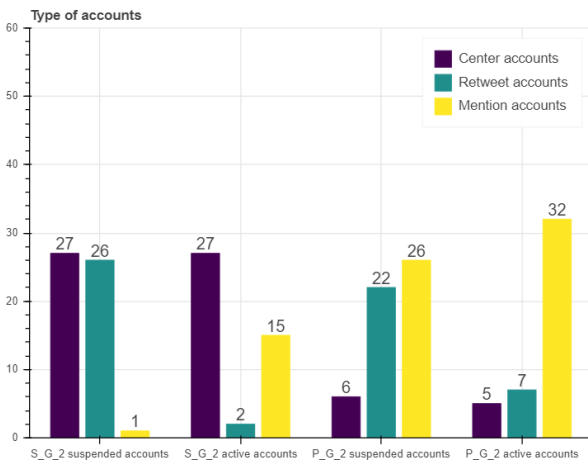


Fig. 12. Type of Accounts that were Suspended or Still Active for the Two Campaigns.

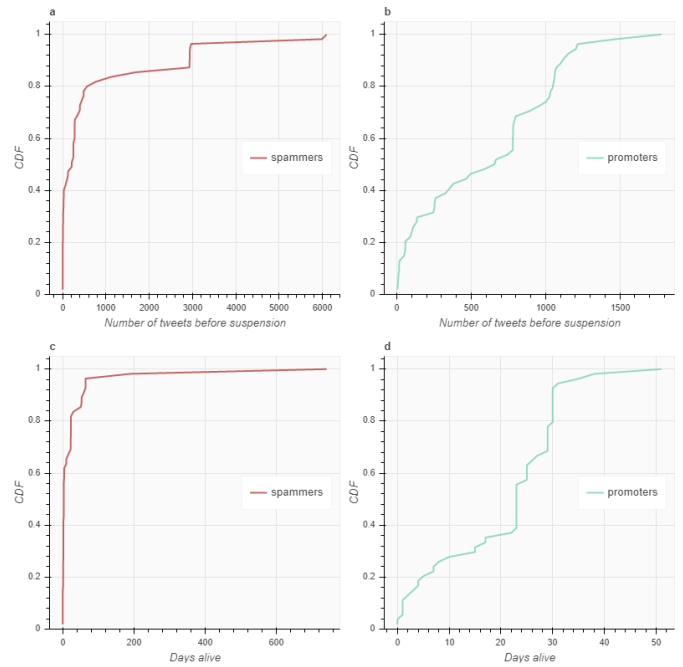


Fig. 13. Lifespans of Suspended Accounts in Days and Total Numbers of Tweets before the Suspension.

C. Practice of Managing Spam Accounts

Malicious campaigns are commonly known to use software to efficiently and quickly manage an enormous number of fake accounts [14], [40], [41]. Such accounts are known as botnets, which are fake accounts on social media that are fully or partially controlled by software. A synchronized or identical timestamp is an essential feature that defines such users. In this section, therefore, the timing of activities of the campaigns' accounts are examined to identify botnet-like behavior.

In the case of our dataset, different spamming strategies involve various social interactions such as tweeting, retweeting, mentioning, and favoriting. Accounts that share a high volume of tweets in a short time are more likely to be part of a botnet [36], [42], [43], [44]. Among the two groups, the spammers' accounts tended to post a larger number of tweets daily (see Table III for average and maximum numbers of daily tweets). Secondly, synchronized retweets from a set of accounts are a strong indicator of software or botnet accounts. According to our previous experiment regarding several clusters' timestamps from the two campaigns [12], we found that the accounts have synchronized retweet timestamps. Figure 14 shows an example of a central account's activities over a couple of hours and the retweet timestamps. The account received all the retweets in seconds after posting the original tweets, as is shown in the figure. Thereby, we concluded that these accounts are simultaneously controlled to automatically perform specific tasks such as retweeting, replying, and favoriting. However, we found instances of accounts belonging to the promoters' group that exhibited human-like behavior in which they engaged in meaningful conversation with genuine accounts, for example, answering questions.

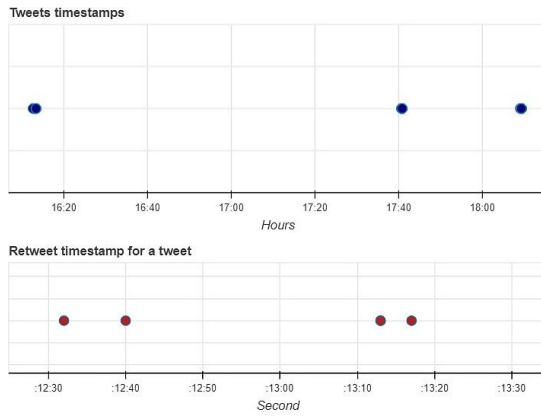


Fig. 14. Timestamps of Original Tweets and Retweets.

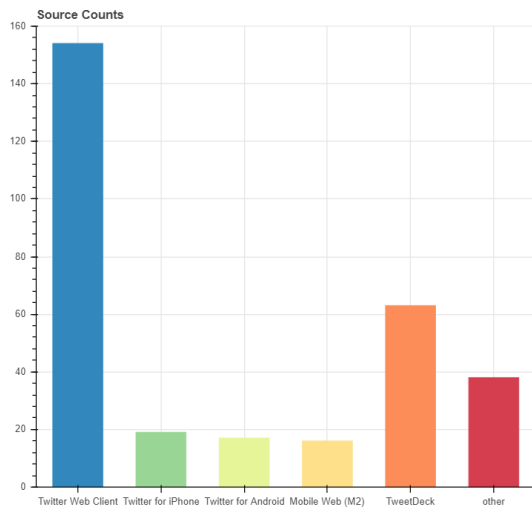


Fig. 15. Sources most used by Spammer Accounts.

The source of an account's tweets might also indicate a sign of possible automated control, wherein some of these sources facilitate this process more than others [45]. We identified two major sources used by the two campaigns: Twitter Web Client and TweetDeck, as illustrated in Figures 15, 16. These two sources are generally known for services that involvescheduling future activities and managing multiple accounts. Also, we found that Twitter for iPhone and Twitter for Android were the sources most used by genuine accounts. Additionally, about 40% of the spammer accounts and 38% of the promoter accounts used two sources in their tweets, whereas less than 18% of genuine accounts used two sources. This suggests that some of these accounts are controlled by both humans and software, which results in human- and botnet-like behavior at the same time.

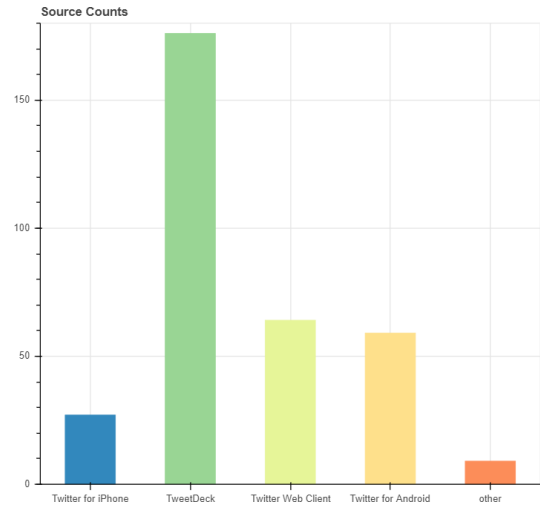


Fig. 16. Sources most used by Promoter Accounts.

V. SUMMARY OF ARABIC SPAM CAMPAIGNS' CHARACTERISTICS

The following points summarize the findings of our in-depth study:

- The spammers' and promoters' campaigns mainly involve accounts that are compromised or stolen from legitimate owners, and their lifespans are often greater than 1,500 days, or four years.
- Both campaigns targeted trending topics through coordinated account groups, and they attempted to manipulate the reputations of their tweets in order to reach to the "top tweets" tab.
- Both campaigns had high self-similarity ratios, with 40% of accounts having a 0.7 similarity ratio, and in general, their content was significantly different from real users' content.
- The spammers' campaign exhibited very easily detectable behavior, as a result of which 60% of accounts were suspended in less than five days. The promoters' accounts, in contrast, posted a fair number of tweets daily, and consequently, their activities (including tweets that were still public in the trends) persisted for more extended periods than those of the spammers.
- We found that the campaigns' accounts were more successful in avoiding Twitter suspension than previously reported in the literature.
- These accounts are partially controlled by human and software, which results in human- and botnet-like behavior at the same time.

VI. CONCLUSION

This paper presents an in-depth analysis of two malicious Arabic-language campaigns on Twitter. They were selected due to illegal practices that were longer-running and more frequent compared to other groups. The primary purpose of this analysis is to examine the content and behavior of malicious Arabic-language groups, including their respective numbers of accounts, spamming tactics, lifespans, and methods used to control these accounts. To this end, we examined six hundred profiles of the two campaigns that were divided into two generations; the first generation's activity took place on Apr-Dec 2018, and the second-generation's activity on Aug-Sep 2019. Through this study, we have shown that compromised accounts that were usually over 1, 500 days or four years old were the accounts most used by the two campaigns.

Both campaigns focused on trends through organized account groups, through which they tried to manipulate their tweets' reputations to reach the top tweets tab; this spamming tactic is clearly shown in their interaction graphs. Secondly, they had straightforward detectable content and their profiles had high ratios of text similarity, with 40% of the accounts having similarity ratios of over 0.7, and in addition, most of them used the same word sets to deliver their messages. Furthermore, we have demonstrated through our 2-month experiment on second-generation accounts that these accounts have avoided Twitter suspension more effectively than has been previously reported in the literature. Among the spammer and promoter campaigns, the promoter campaign was more successful in leveraging their accounts, and their profiles could avoid suspension for a longer period than the spammers' accounts. Finally, they are more likely to use script or software for managing and automating some of their actions, in particular, retweeting and responding.


The analysis provided in this paper has essentially revolved around two malicious Arabic-language campaigns on Twitter. Although most other malicious campaigns either disseminate spam URLs or promote a certain product or service, some of these groups are worth investigating in future research. For example, we found through this study that many groups of fake accounts that offer a service are explicitly manipulating trending topics. This is an interesting area for future work: first to study the trending topics and identify low-quality trending hashtags or topics, and second to investigate the techniques used by these accounts to manipulate topics.

REFERENCES

- [1] N. Eltantawy and J. B. Wiest, "The arab spring— social media in the egyptian revolution: reconsidering resource mobilization theory," *International journal of communication*, vol. 5, p. 18, 2011.
- [2] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad, "Opening closed regimes: what was the role of social media during the arab spring?" *Available at SSRN 2595096*, 2011.
- [3] A. Bruns, T. Highfield, and J. Burgess, "The arab spring and social media audiences: English and arabic twitter users and their networks," *American behavioral scientist*, vol. 57, no. 7, pp. 871–898, 2013.
- [4] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar, "Opinion mining and analysis for arabic language," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181–195, 2014.
- [5] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? disruptive event detection using twitter," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 2, pp. 1–26, 2017.
- [6] H. Almerkhi, M. Hasanain, and T. Elsayed, "Evetar: A new test collection for event detection in arabic tweets," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 689–692.
- [7] S. M. Alzanin and A. M. Azmi, "Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowledge-Based Systems*, vol. 185, p. 104945, 2019.
- [8] G. Wolfsfeld, E. Segev, and T. Sheaffer, "Social media and the arab spring: Politics comes first," *The International Journal of Press/Politics*, vol. 18, no. 2, pp. 115–137, 2013.
- [9] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: Dark of the social networks," *Journal of Network and Computer Applications*, vol. 79, pp. 41–67, 2017.
- [10] H. S. Al-Khalifa, "On the analysis of twitter spam accounts in saudi arabia," *International Journal of Technology Diffusion (IJTD)*, vol. 6, no. 1, pp. 46–60, 2015.
- [11] M. Alfifi and J. Caverlee, "Badly evolved? exploring long-surviving suspicious users on twitter," in *International Conference on Social Informatics*. Springer, 2017, pp. 218–233.
- [12] R. Alharthi, A. Alhothali, and K. Moria, "Detecting and characterizing arab spammers campaigns in twitter," *Procedia Computer Science*, vol. 163, pp. 248–256, 2019.
- [13] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 963–972.
- [14] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copy-catch: stopping group attacks by spotting lockstep behavior in social networks," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 119–130.
- [15] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali, "Nd-sync: Detecting synchronized fraud activities," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 201–214.
- [16] A. Duh, M. Slak Rupnik, and D. Koros'ak, "Collective behavior of social bots is encoded in their temporal twitter activity," *Big data*, vol. 6, no. 2, pp. 113–123, 2018.
- [17] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.
- [18] N. Vo, K. Lee, C. Cao, T. Tran, and H. Choi, "Revealing and detecting malicious retweeter groups," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2017, pp. 363–368.
- [19] H. AlMahmoud and S. AlKhalifa, "Tsim: a system for discovering similar users on twitter," *Journal of Big Data*, vol. 5, no. 1, p. 39, 2018.
- [20] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 71–80.
- [21] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benvenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 61–70.
- [22] S. Gupta, D. Kuchhal, P. Gupta, M. Ahamad, M. Gupta, and P. Kumaraguru, "Under the shadow of sunshine: Characterizing spam campaigns abusing phone numbers across online social networks," in *Proceedings of the 10th ACM Conference on Web Science*, 2018, pp. 67–76.
- [23] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Computers & Security*, vol. 76, pp. 265–284, 2018.
- [24] N. El-Mawass and S. Alaboodi, "Detecting arabic spammers and content polluters on twitter," in *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*. IEEE, 2016.

- pp. 53–58.
- [25] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on twitter,” in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.
- [26] N. Al Twaresh, M. Al Tuwajri, A. Al Moammar, and S. Al Humoud, “Arabic spam detection in twitter,” in *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, 2016, p. 38.
- [27] H. Almerexhi and T. Elsayed, “Detecting automatically-generated arabic tweets,” in *AIRS*. Springer, 2015, pp. 123–134.
- [28] Z. Chen and D. Subramanian, “An unsupervised approach to detect spam campaigns that use botnets on twitter,” *arXiv preprint arXiv:1804.05232*, 2018.
- [29] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, “Detecting and characterizing social spam campaigns,” in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 35–47.
- [30] Z. Alom, B. Carminati, and E. Ferrari, “Detecting spam accounts on twitter,” in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 1191–1198.
- [31] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh, A. Alarifi, A. Alfariis *et al.*, “If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts,” *International Journal of Information Security*, vol. 15, no. 5, pp. 475–491, 2016.
- [32] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, “Compa: Detecting compromised accounts on social networks.” in *NDSS*, 2013.
- [33] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, “Profiling online social behaviors for compromised account detection,” *IEEE transactions on information forensics and security*, vol. 11, no. 1, pp. 176–187, 2015.
- [34] P. Bindu, R. Mishra, and P. S. Thilagam, “Discovering spammer communities in twitter,” *Journal of Intelligent Information Systems*, vol. 51, no. 3, pp. 503–527, 2018.
- [35] A. A. Amleshwaram, A. N. Reddy, S. Yadav, G. Gu, and C. Yang, “Cats: Characterizing automation of twitter spammers.” in *COMSNETS*, 2013, pp. 1–10.
- [36] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization. corr abs/1703.03107 (2017),” *arXiv preprint arXiv:1703.03107*, vol. 3, 2017.
- [37] E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Danforth, and P. S. Dodds, “Sifting robotic from organic text: a natural language approach for detecting automation on twitter,” *Journal of computational science*, vol. 16, pp. 1–7, 2016.
- [38] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, “Aravec: A set of arabic word embedding models for use in arabic nlp,” *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.
- [39] R. Alharthi, A. Alhothali, and K. Moria, “A real-time deep-learning approach for filtering arabic low-quality content and accounts on twitter,” *Information Systems*, vol. 99, p. 101740, 2021.
- [40] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.
- [41] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling,” *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 561–576, 2017.
- [42] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, “A new approach to bot detection: striking the balance between precision and recall,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 533–540.
- [43] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, “Botornot: A system to evaluate social bots,” in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 273–274.
- [44] A. H. Wang, “Detecting spam bots in online social networking sites: a machine learning approach,” in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2010, pp. 335–342.
- [45] C. Yang, R. Harkreader, and G. Gu, “Empirical evaluation and new design for fighting evolving twitter spammers,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

Integrating Cost-231 Multiwall Propagation and Adaptive Data Rate Method for Access Point Placement Recommendation

Fransiska Sisilia Mukti¹, Puput Dani Prasetyo Adi^{2*}, Dwi Arman Prasetya³
Volvo Sihombing⁴, Nicodemus Rahanra⁵, Kristia Yuliawan⁶, Julianto Simatupang⁷

Department of Informatic, Technology and Design Faculty, Institut Teknologi dan Bisnis Asia Malang, Indonesia¹

Department of Electrical Engineering, Engineering Faculty, University of Merdeka Malang^{2,3}

Department of Informatics Management, Faculty of Science and Technology, Universitas Labuhanbatu⁴

Department of Informatics Engineering, Faculty of technology and Engineering, Universitas Satya Wiyata Mandala⁵

Department of Informatics Engineering, Engineering Faculty, Universitas Negeri Papua⁶

Department of Information Technology, AMIK Mahaputra, Riau⁷

Abstract—A new approach has been developed to provide an overview about signal behavior in indoor environments using Cost-231 Multiwall Model (Cost-231 MWM) and Adaptive Data Rate (ADR) method. This approach used as a reference for access point (AP) placement for campus building. The Cost-231 MWM plays a role in estimating the measured power received by user (usually called as Received Signal Strength Indicator/RSSI) by considering the existence of obstacles around the transmitter (AP). We used Institut Asia Malang environments as the case study and gave some recommendations for AP placement: ten optimal placements for the first, third and fourth floor, also seven optimal placements for the second floor. These recommendations were based on the RSSI for good and excellent level signal (-50 dBm until -10dBm). This research also uses the Adaptive Data Rate (ADR) mechanism approach to reduce the amount of packet loss (kbps) resulting from obstacles that cause attenuation (-dB). With the Adaptive Data Rate mechanism, it means increasing the number of access points, the signal attenuation (-dB) occurs from the obstacles (Walls) that are penetrated by the Radio Frequency device and causes attenuation (-dB), the more Access points on the Multi-Wall, will allow communication and data transmitting stability.

Keywords—Access point placement; indoor propagation; Cost-231 Multiwall; ADR; RSSI

I. INTRODUCTION

Recently, wireless network technology is not something foreign to society but has become one of the main communication media infrastructures over time. Refers to the IEEE 802.11, wireless network has been developing rapidly [1], [14],[18]. This technology uses electromagnetic waves for communication between nodes. Flexibility and mobility were the main points why this technology is in great demand by users than cable line [2]. For the local area, we usually named this technology as Wireless Local Area Network (WLAN). WLAN is becoming familiar wireless technology built as an extension of a wired LAN [3], [24].

Even though it provides convenience in the installation and configuration process, building a wireless network cannot be underestimated. A network engineer must understand the

environments clearly about the interferences that may occur on-site, especially for indoor environments. This factor makes the placement of wireless transmitter devices (named as Access Point or AP) quite tricky to do because misplacement of AP will cause decreasing in communication performance.

Specifically, understanding the propagation of a signal from transmitter to receiver in wireless communication is studied in propagation. Indoor and outdoor propagation have different parameters; even indoor propagation provides more complicated parameters than outdoor propagation [4]. This is due to the presence of the materials around the APs could attenuate the signal while transmitted to the user [2], such as reflection, diffraction, or scattering [5], [11], [13], [16].

Studies were conducted to find out the best approach for indoor propagation, [6] both mathematical equations and based on site-survey measurement. Some new models were also developed to evaluate the signal's behaviors for various environments. Different propagation models also presented the effects of building layout and found the best approach for the environment. [7] Statistical models have been considered as an excellent strategy in designing wireless infrastructure without the need for detailed analysis indeed.

Several studies were deployed to evaluate the accuracy of empirical propagation for the indoor environment. the Cost-231 Multiwall propagation model and the Offered Bit Quantity method to determine the optimal number of APs and the placement. The results show that this method provides a better coverage area, and a more substantial signal strength value reaches -27.27 dBm [8]. An APs placement design must consider the importance of propagation losses. The calculation used empirical propagation based on the areas [9]. Mukti was compared four types of propagation : one slope, log-distance, cost-231 MWM and ITU-R, to figure out which the most suitable modelling for campus environment. For that case, ITU-R (P.1238) model gave the closest results to actual measurement with 16,381% relative error rate [5].

As one of the educational institution, Institut Asia Malang used WLAN as its wireless infrastructure. This place consists

*Corresponding Author

of four floors where each floor has several APs. Previous research [6] showed that there are still several locations categorized as blank spots (with the poor signal level of 0-39%) because AP placement is only based on the officer's feeling, without considering the aspect of propagation. Based on this case, this study aims to bring an idea in AP's placement using Cost-231 Multiwall Model (MWM) as one of the empirical propagation models and integrate it with adaptive rate method for improving the received signal level for the users. This approach will take a concern about environmental interferences such as floors, walls, doors, etc. The results will be used as a recommendation for the related part of the institution.

II. THEORY

A. COST-231 Multi-Wall Model

Cost-231 MWM gives better accuracy than the earlier model such One Slope Model (1SM), because it used environment description as the input variable [3]. Overview about this model shown in Figure 1, while the pathloss value between transmitter and receiver calculated using Equation 1.

$$L_{MW} = L_{FSL}(d) + \sum(N, i = 1) k_{wi}L_{wi} + k_f L_f \quad (1)$$

It is important to pay attention of the parameters such as wall attenuation, to get the closest prediction. L_{wi} does not represent actual value, but only a statistical value from representative calculations in previous studies. There are two types of wall in Cost-231 MWM: light wall (L1) with thin wall or partitions, and heavy walls (L2) as thick structured walls. Meanwhile, other parameters was defined in Table 1 and Table 2 [3][7].

B. Free Space Path Loss (FSPL)

FSPL defined as losses of RF signal while reaching certain distance (between transmitter and receiver antenna). Equation 2 specifically gives mathematical equation to find this value [3][5].

$$L_{FSL}(d) = 32,44 + 20 \log (d) + 20 \log (f) \quad (2)$$

where L_{FSL} as free space loss value in dB, d refers to TX and RX distance in meters and f as AP's frequency in MHz.

C. Received Signal Strength Indicator (RSSI)

In telecommunications, received signal strength indicator (RSSI) is a measurement of the power present in a received radio signal. A well-managed wireless network can provide a good RSSI value (a negative value is said to be 0 dB) [4][3][5][28]. The RSSI value is obtained from Equation 3.

$$RSSI = EIRP - F_{SL} + G_R - L_{MW} \quad (3)$$

where RSSI inform the received signal by user in dB, $EIRP$ as the power of AP when transmit data in dBm, G_R as RX gain in dBm and L_{MW} as the pathloss value calculation using Cost-231 MWM.

D. Effective Isotropic Radiated Power (EIRP)

EIRP is the total energy expended by an access point and antenna. When an access point sends its energy to the antenna to be transmitted, a large reduction in energy will occur in the

cable. To compensate for this, an antenna adds power / gain, the amount of additional power will depend on the type of antenna used [10]. Equation 4 represent the calculation of this value.

$$EIRP = P_T + G_R - L_{MW} \quad (4)$$

where P_T as the transmit power of the device in dBm.

E. Data Rate and Sensitivity

Data rate also determined as spreading factor, are influenced by the distance between TX and RX. The farther the distance, the weaker the signal strength. It will affects the throughput of the communication. The throughput is getting smaller and the packet loss is getting bigger. And the greater the value of the spreading factor, the longer it takes the transmitter to reach the receivers or Time on Air (ms) [11], [22], [24].

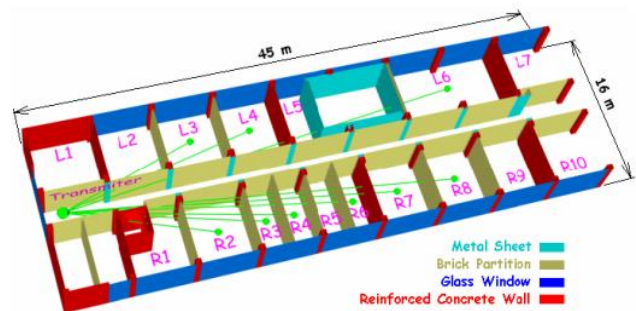


Fig. 1. Geometries Multi-Wall Model.

TABLE I. ATTENUATION VALUE OF RADIO FREQUENCY ON THE INDOOR BUILDING MATERIALS

Nu.	Parameter	Attenuation Value (dB)
1.	Cubicle wall	2
2.	Wooden door	3
3.	Glass window	3
4.	Drywall or sheetrock	3
5.	Metal shelf	6
6.	Elevator or metal particles	10
7.	Brick, concrete, concrete block	12
8.	Ceramic Floor	13.2
9.	Foundation wall	15

TABLE II. SIGNAL ATTENUATION VALUE AT 2.4 GHZ FREQUENCY

Nu.	Parameter	Attenuation Value (dB)
1.	Brick wall window	2
2.	Brick wall next to metal door	3
3.	Cinder block wall	4
4.	Office walls	6
5.	Office wall metal doors	6
6.	Metal glass wall frame	6
7.	Metal door on brick wall	12.4

TABLE III. DATA RATE, SENSITIVITY AND TIME ON AIR

Data Rate (Spreading Factor)	Time on Air	Sensitivity
SF7	41 ms	-123 dBm
SF8	72 ms	-126 dBm
SF9	144 ms	-129 dBm
SF10	288 ms	-132 dBm
SF11	577 ms	-134.5 dBm
SF12	991 ms	-137 dBm

III. METHOD USED

A. Real Measurement and Analyze use inSSIDer

This research is a continuation from some previous studies [3][6]. Located in Institut Asia Malang, this research used the site survey measurements which calculated repeatedly using regression method and inSSIDER application to get real the signal strength values. The measurement was built in two propagation paths: Line of Sight (LoS) path and Non Line of Sight (NLoS) path. More than 25 measurement points were taken for each floor in order to get the best accuracy on sampling (close to 90%).

To find out the closest RSSI level prediction compared to the real values from site survey measurements before, we evaluated all propagation parameters and calculated it into Equation 1 to 4 with detail specifications below (Tabel IV) [3].

We analyzed every points and elected the optimal placement based on best RSSI level, both on LoS and NLoS path, for excellent signal level (see table V) [21],[25],[26],[27].

B. Adaptive Data Rate (ADR) Mechanism Approach

Adaptive Data Rate (ADR) [12],[16] is a mechanism for increasing the number of receivers or access points at a certain point which aims to amplify the signal transmitted by the transmitter (Tx) in an internet network [19]. The additional AP is then sent to another AP in a condition that it is blocked by a wall of different thickness, type, and wall material. This affects the state of the signal resulting in an attenuation signal [17], [22].

Therefore, the function of the Additional AP is to strengthen signal reception in receivers or EDs and reduce packet loss or increase throughput [20]. Figure 2 is an example of ADR representation to make it easier to understand Adaptive Data Rate Mechanism on Multi Wall.

The Adaptive Data Rate Schedule mechanism in Figure 3 consists of three critical parameters, i.e., Uplink, Downlink, and ADR Response. In the Uplink Process, the ADR must be determined the data bits will be sent; therefore, they are recorded in the uplink data process using the ADR Ack bit. Some ADR parameters originating from the downlink, i.e., ADR scheduled, ADR failed, Collected data, and ADR is running. The last thing is about ADR Response, if successful then ADR Success, and go to the un-schedule ADR process. The ADR algorithm is often used for Low Data Rate data such as that of LoRa and together with Spreading Factor analysis (6-12) [15],[20], [23].

TABLE IV. PROPAGATION PARAMETER SPECIFICATIONS

Nu.	Parameter	Value
1.	Operating band (frequency / f)	2.4 GHz
2.	TX gain	3 dBi
3.	RX gain (G _R)	0 dBi
4.	Maximum TX Power	27 dBm
5.	Line losses	0.5 dB
6.	Fading margin of WLAN	10 dB

TABLE V. RSSI CATEGORY FOR WLAN

Category	Range (dBm)	Percentage (%)
Excellent	-57 to -10	75 – 100
Good	-75 to -58	40 – 74
Fair	-85 to -76	20 – 39
Poor	-95 to -86	0 – 19

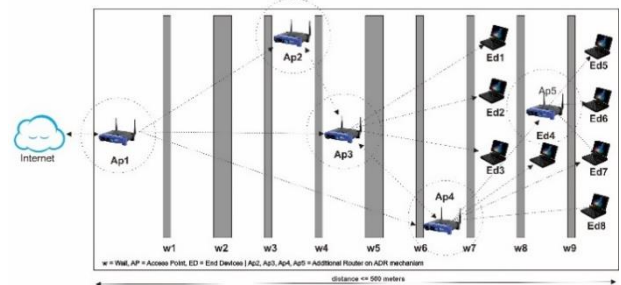


Fig. 2. Adaptive Data Rate Mechanism on Multi Wall.

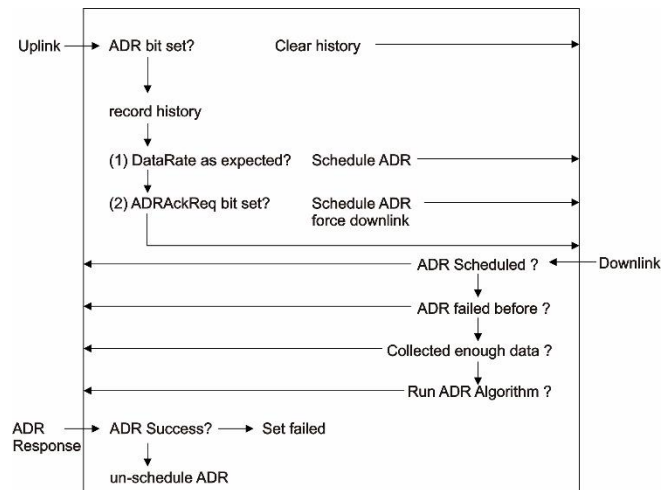


Fig. 3. Adaptive Data Rate Schedule Mechanism.

IV. RESULT AND DISCUSSION

The analysis started with evaluating the building structure for each floor. First floor showed more free space area, while the second, third and fourth floor seemed to be closed building with some rooms and corridors. We made some prediction location for AP placement and adopted 12 measurement points in first floor, both indoor and outdoor placement. We also defined every obstacle around the APs so that we can calculating the values, as shown in Figure 4.



Fig. 4. Floor Plan 1 along with Obstacle Description.

Because every obstacle has their own attenuation values (see Table II and Table III), we must identify it exactly, to avoid miscalculated prediction. These identification process will be used to find the pathloss and EIRP value for every placement point using Cost-231 MWM and ADR method. The sample results shown at Table VI. Propagation theory for closed room claimed that the best transmitted signal occurred while in LoS propagation (there are no barrier between AP and user) [6], [11], [12]. We could see the LoS path for the first floor are placement on number 1, 3, 5, 8, 9 and 12. The best signal coverage is reached for indoor only, while the outdoor area (NLoS path) could not capture the signal properly. Based on these calculations, we took only -58 dBm to -10 dBm values for recommendation placements (defined as excellent and good signal levels, see Table V). Therefore, we recommend number 1 until 10 as the placement points.

TABLE VI. EIRP CALCULATION FOR AP PLACEMENT POINT SAMPLES ON 1ST FLOOR

AP	Obstacles		EIRP (dB)
	Type	Attenuation dB)	
1	Ceramic Floor	13.2	4.8
	Glass	3	
	Glass door	6	
	Wood Dividers	3	
2	Wall	6	12
	Wood cupboard	3	
	Glass window	3	
3	Wooden door	3	1.8
	Glass	3	
	Ceramic Floor	13.2	
	Wooden partition	3	
4	Wooden door	3	1.8
	Wall	6	
	Glass table	6	
	Ceramic Floor	13.2	
5	Ceramic Floor	13.2	-1.2
	Glass	3	
	Metal frame	6	
	Wooden door	3	
	Wall	6	

The same process was carried out for second, third and fourth floors. Even though all of them have the same building structure, however, we still carry out an in-depth analysis for each floor and found 7 optimal placements for second floor, 10 optimal placements for third floor and 9 optimal placements for fourth floor. Furthermore, we figured out the RSSI values prediction for each placement and compared it with our previous studies (site-survey measurements and One Slope Model). In order to obtain the accuracy validity of the comparison, we used the same test point.

In order to get an overview of the comparison of the proposed methods, we visualize the results of our observations through Figure 5-8. For the first floor, we used placement point number 12 and calculated the RSSI values for each approach (see Figure 5). Meanwhile in second floor, there is only one AP placed in center of the corridor, and the results showed on Figure 6.

3rd floor became most crowded place because this place consists of 9 classrooms and lecturer’s room. Almost all of the lecture activities are carried out in this area. Two AP are placed in this area: center of the corridor and in the lecture room. We compared the results into a graph on Figure 7. Hereafter, we got some calculation also for the fourth floor, and showed the analysis into a graph on Figure 8.

Power Level Comparison (-dBm) with ADR Approach on 1st floor

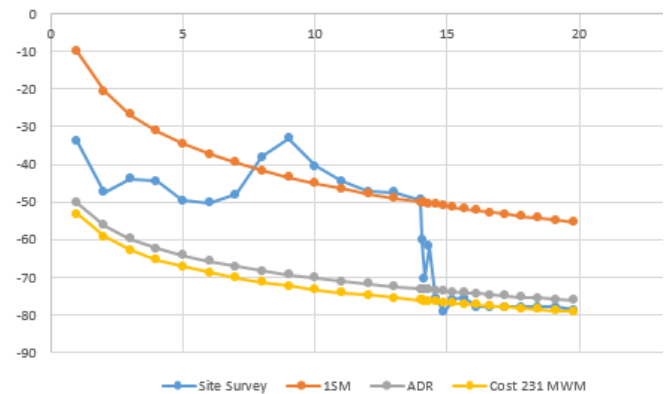


Fig. 5. Power Level Comparison on 1st Floor.

Power Level Comparison (-dBm) with ADR Approach on 2nd floor

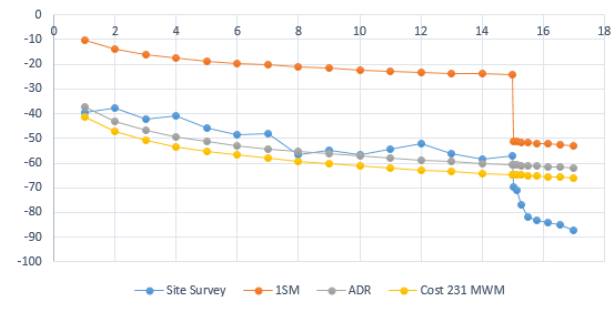


Fig. 6. Power Level Comparison on 2nd Floor.

Power Level Comparison (-dBm) with ADR Approach on 3rd floor

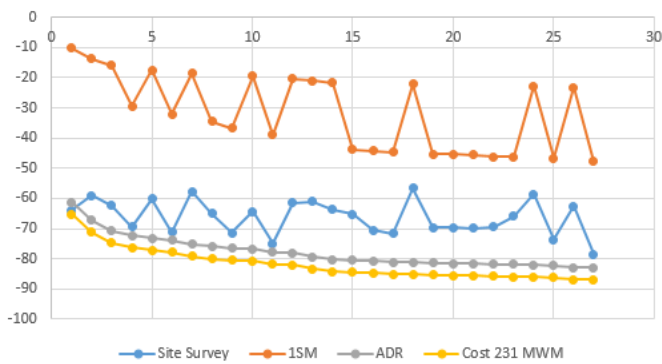


Fig. 7. Power Level Comparison on 3rd Floor.

Power Level Comparison (-dBm) with ADR Approach on 4th floor

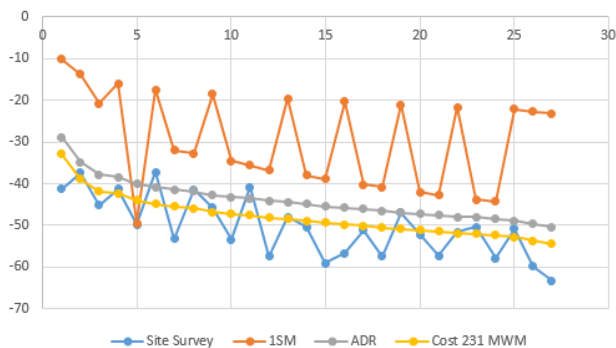


Fig. 8. Power Level Comparison on 4th Floor.

The comparison of power levels shown in Figure 5 to Figure 8 shows that Cost-231 MWM and ADR method give closest value than One Slope Mode on previous studies. It proves that our approach can be used as RSSI prediction method for indoor environments, because we considered the existence of the obstacles which has significant effect for the RSSI level. Therefore, it is essential to pay attention for AP placement.

V. CONCLUSIONS

Our research was built to find the closest prediction in RSSI level for indoor environments. It is greatly influenced by the existence of the obstacles between AP and user. The Cost-231 MWM approach provides a closest propagation values which compared with actual values based on site-survey measurement. Our analysis proved that the obstacle gives significant impact for the user's signal level (RSSI). Signal strength analysis was performed on the LOS and NLOS propagation paths. We showed there were 10 optimal placements for first, third and fourth floor, also 7 optimal placements for second floor. These recommendation was choose based on signal strength susceptibility on -58dBm to -10dBm for LoS propagation.

Our approach can be used as reference for the related division on Institut Asia Malang in reviewing the current AP placement. The objective of this research to provide a better area coverage and WLAN performance for case study. Further, the ADR method helps stabilize data from the transmitter and reduces data loss due to the large number of walls which results in attenuation of the signal from the transmitter and reduces the load on the Access point (APs).

A. Future Work

The analysis will be improved by increasing the number of APs using ADR methods and measuring the multistoried buildings and analyzing the signal measurement between multi-storey buildings and the number of houses in one area. Devices will be developed not only using WiFi, however, a combination of WSNs and LPWAN devices e.g., LoRaWAN and additional analysis using software and hardware to analyze signal strength or Radio Frequency in realtime. And as the additional analyzes is a Spreading Factor analyzes on the measurement. Also, this research can be developed into a desktop or android system which has ability to give an real-time overview for signal distribution.

ACKNOWLEDGMENT

Thanks to Institut Asia Malang, which has become a place for research and data collection so that this research can be completed properly, and thanks to the entire academic community of Institute Asia Malang, and University of Merdeka Malang (UNMER-Malang).

REFERENCES

- [1] A. T. Suryani and A. B. Pantjawati, "Analysis of the Coverage Area of the Access Point Using Netspot Simulation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 384, no. 1, 2018, doi: 10.1088/1757-899X/384/1/012001.
- [2] F. S. Mukti and A. Junikhah, "A Coverage Prediction Technique for Indoor Wireless Campus Network," *J. INFOTEL*, vol. 11, no. 3, pp. 73–79, 2019, doi: 10.20895/infotel.v11i3.434.
- [3] F. S. Mukti, "Access Point Placement Recommendation Using Cost-231 Multiwall Propagation," *J. Penelit. Pos dan Inform.*, vol. 10, no. 2, p. 103, 2020, doi: 10.17933/jppi.2020.100202.
- [4] F. S. Mukti and A. Junikhah, "Prediksi Cakupan Area untuk Jaringan Wireless Indoor Kampus berdasarkan Penempatan Access Point," *J. Penelit. Ilmu Tek. dan Terap.*, vol. 10, no. 2, pp. 67–72, 2019, doi: https://doi.org/10.48056/jintake.v10i2.55.
- [5] F. S. Mukti, "Studi Komparatif Empat Model Propagasi Empiris Dalam Ruang untuk Jaringan Nirkabel Kampus," *J. Teknol. dan Sist. Komput.*, vol. 7, no. 4, pp. 154–160, 2019, doi: 10.14710/jtsiskom.7.4.2019.154-160.
- [6] F. S. Mukti and D. A. Sulisty, "Analisis Penempatan Access Point Pada Jaringan Wireless Lan Stmik Asia Malang Menggunakan One Slope Model," *J. Ilm. Teknol. Inf. Asia*, vol. 13, no. 1, pp. 13–22, 2018, doi: 10.32815/jitika.v13i1.304.
- [7] F. J. Carlos Vesga, H. Martha Fabiola Contreras, and B. Jose Antonio Vesga, "Design of empirical propagation models supported in the Log-Normal Shadowing model for the 2.4GHz and 5GHz bands under Indoor environments," *Indian J. Sci. Technol.*, vol. 11, no. 22, pp. 1–18, 2018, doi: 10.17485/ijst/2018/v11i22/122149.
- [8] M. A. Amanaf, E. S. Nugraha, and L. Azhari, "Analisis Optimasi Perencanaan Ulang Access Point Wifi Dengan Model Pathloss COST 231 Multi Wall dan Metode Offered Bit Quantity (OBQ) Studi Kasus Gedung Telematika ITTP," *J. Telecommun. Electron. Control Eng.*, vol. 1, no. 01, pp. 32–42, 2019, doi: 10.20895/jtece.v1i01.39.
- [9] P. Titahningsih, R. Pramananda, and S. R. Akbar, "Perancangan Penempatan Access Point untuk Jaringan Wifi Pada Kereta Api

- Penumpang,” J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 5, pp. 2008–2015, 2018.
- [10] A. Fauzi and M. Arrofiq, “Assesment Kekuatan Daya Received Signal Level (RSL) Wireless 2.4GHz di Ruang meeting,” J. Internet Softw. Eng., vol. 1, no. 1, pp. 10–17, 2020.
- [11] Adi, P.D.P, Kitagawa, "A performance of radio frequency and signal strength of LoRa with BME280 sensor", *Telkonnika (Telecommunication Computing Electronics and Control)*, Issue 2, 1 April 2020, Pages 649-660, DOI: 10.12928/telkonnika.v18i2.14843.
- [12] Kunho Park, Junhyun Park, Junhyun Park, A-Hyun Lee, Chong-kwon Kim, "An Energy Efficient ADR Mechanism Considering Collision Rate in LoRa Network", December 2020, *KIISE Transactions on Computing Practices* 26(12):535-540, DOI: 10.5626/KTCP.2020.26.12.535.
- [13] Puput Dani Prasetyo Adi and Akio Kitagawa, "ZigBee Radio Frequency (RF) Performance on Raspberry Pi 3 for Internet of Things (IoT) based Blood Pressure Sensors Monitoring" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 10(5), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0100504>.
- [14] Fabian Rincon, et.al., "On the Impact of WiFi on 2.4 GHz Industrial IoT Networks", October 2018, Conference: 2018 IEEE International Conference on Industrial Internet (ICII), DOI: 10.1109/ICII.2018.00012.
- [15] Norhane Benkahla, Hajer Tounsi, Ye-Qiong Song, Mounir Frikha, "Review and experimental evaluation of ADR enhancements for LoRaWAN networks", January 2021, Springer, *Telecommunication Systems*, DOI: 10.1007/s11235-020-00738-x.
- [16] Puput Dani Prasetyo Adi and Akio Kitagawa, "Quality of Service and Power Consumption Optimization on the IEEE 802.15.4 Pulse Sensor Node based on Internet of Things" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 10(5), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0100518>.
- [17] Derek Heeger, et.al, "Secure LoRa Firmware Update with Adaptive Data Rate Techniques", March 2021, *Sensors* 21(7):2384, DOI: 10.3390/s21072384.
- [18] Aathmanesan T, "Design of Metamaterial Antenna for 2.4 GHz WiFi Applications", August 2020 *Wireless Personal Communications* 111(4), Springer, DOI: 10.1007/s11277-020-07324-z.
- [19] Puput Dani Prasetyo Adi and Akio Kitagawa, "Performance Evaluation of E32 Long Range Radio Frequency 915 MHz based on Internet of Things and Micro Sensors Data" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 10(11), 2019. DOI: 10.14569/IJACSA.2019.010110.
- [20] Smangalis Mnguni, "Performance Evaluation of Spreading Factors in LoRa Networks", In book: *Towards new e-Infrastructure and e-Services for Developing Countries*, 12th EAI International Conference, AFRICOMM 2020, Ebène City, Mauritius, December 2-4, 2020, Proceedings, DOI:10.1007/978-3-030-70572-5_13.
- [21] Stancu Eugen, et.al, "Spectral Analysis in the 2.4 GHz WiFi Band in Bucharest", June 2020, Conference: 2020 13th International Conference on Communications (COMM), DOI: 10.1109/COMM48946.2020.9142040.
- [22] Puput Dani Prasetyo Adi and Akio Kitagawa, "A Study of LoRa Performance in Monitoring of Patient's SPO2 and Heart Rate based IoT" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 11(2), 2020. doi. 10.14569/IJACSA.2020.0110232.
- [23] Jaber Babaki, et.al, "Dynamic Spreading Factor and Power Allocation of LoRa Networks for Dense IoT Deployments", Conference: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, August 2020, DOI: 10.1109/PIMRC48278.2020.9217283.
- [24] Puput Dani Prasetyo Adi and Akio Kitagawa, "Performance Evaluation WPAN of RN-42 Bluetooth based (802.15.1) for Sending the Multi-Sensor LM35 Data Temperature and RaspBerry Pi 3 Model B for the Database and Internet Gateway" *International Journal of Advanced Computer Science and Applications(ijacs)*, 9(12), 2018. <http://dx.doi.org/10.14569/IJACSA.2018.091285>.
- [25] Muzaiyanah Hidayab, Abdul H Ali, Khairul Bariah Abas Azmi, "Wifi signal propagation at 2.4 GHz", January 2010, Conference: Microwave Conference, 2009. APMC 2009. Asia Pacific, DOI: 10.1109/APMC.2009.5384182.
- [26] Petrus Kerowe Goran, Eka Setia Nugraha, "Asymmetric-Slit Method on WiFi Antenna with 2.4 GHz and 5 GHz Frequency", September 2020 *IJITEE (International Journal of Information Technology and Electrical Engineering)* 4(2):53, DOI: 10.22146/ijitee.55811.
- [27] Mohamed Ibrahim Alhajri, "2.4 GHZ Indoor Channel Measurement", Project: Machine Learning for Contemporary Communication Systems: Methods & Applications, November 2018, DOI: 10.21227/ggh1-6j32.
- [28] Muhammad Niswar., et.al, "Performance evaluation of ZigBee-based wireless sensor network for monitoring patients' pulse status", Conference: *Information Technology and Electrical Engineering (ICITEE)*, 2013 International, DOI: 10.1109/ICITEE.2013.6676255, Oct 2013.

A Comparative Analysis of Hadoop and Spark Frameworks using Word Count Algorithm

Yassine Benlachimi¹, Abdelaziz El Yazidi², Moulay Lahcen Hasnaoui³
ENSAM Moulay-Ismaïl University, LMMI Laboratory, Institute, Meknes, 50000, Morocco

Abstract—With the advent of the Big Data explosion due to the Information Technology (IT) revolution during the last few decades, the need for processing and analyzing the data at low cost in minimum time has become immensely challenging. The field of Big Data analytics is driven by the demand to process Machine Learning (ML) data, real-time streaming data, and graphics processing. The most efficient solutions to Big Data analysis in a distributed environment are Hadoop and Spark administered by Apache, both these solutions are open-source data management frameworks and they allow to distribute and compute the large datasets across multiple clusters of computing nodes. This paper provides a comprehensive comparison between Apache Hadoop & Apache Spark in terms of efficiency, scalability, security, cost-effectiveness, and other parameters. It describes primary components of Hadoop and Spark frameworks to compare their performance. The major conclusion is that Spark is better in terms of scalability and speed for real-time streaming applications; whereas, Hadoop is more viable for applications dealing with bigger datasets. This case study evaluates the performance of various components of Hadoop—such, MapReduce, and Hadoop Distributed File System (HDFS) by applying it to the well-known Word Count algorithm to ascertain its efficacy in terms of storage and computational time. Subsequently, it also provides an analysis of how Spark's in-line memory processing could reduce the computational time of the Word Count Algorithm.

Keywords—Big data; Hadoop; spark; machine learning; Hadoop Distributed File System (HDFS); MapReduce; word count

I. INTRODUCTION

Due to the advancements in computational technology, hardware resources, and fast underlying networks, the world witnessing an explosion of Big Data generated by social media networks [1], Internet of Things (IoT)[2], streaming real-time applications [3], banking sector [4], industrial setups, and almost every notable R&D sector. According to [5] an estimate by a well-known online source, Social Media Today, 2.5 Exabyte (1018) data is generated per day, as of 2020. This data creation is expected to increase to 463 Exabytes per day by the end of 2025, according to Statista [6]. Consequently, it becomes extremely difficult to handle such enormous volumes of Big Data by using traditional methods and tools [7]. For example, the traditional database systems administering the legacy warehouses have become inefficient due to the utilization of conventional query tools. Venkatraman et al. [8] found multiple reasons for the failure of these tools. Firstly, the design of relational databases and data warehouses is not suitable to synthesize the new types of data with respect to volume, storage, veracity, and processing. Secondly, in traditional systems, the Structured Query Language (SQL) is

utilized for communicating with databases. Thirdly, the maintenance of rational data-houses becomes very costly and unmanageable. Fourthly, traditional warehouses are based on organizing records in fields in a structured manner, while most of the Big Data on the Internet is unstructured by nature [9]. Therefore, traditional database management tools cannot efficiently be utilized in the case of Big Data, which is exponentially growing due to the surge in the number of Internet & social media users, and the development of new technologies like IoT, 5G networks, and Deep Learning (DL), etc. This entails an extremely competitive atmosphere among technology companies to provide accurate data in a minimum amount of time at a low cost. It is the only concept of Big Data that gives equal opportunity to everyone to extract the data and use the full value from in their particular organization or concerned field of interest. Chen et al. and Ward et al. [10, 11] more formally defined Big Data as “a set of several structured, unstructured data generated from different formats of various tasks with bulk volume that is uncontrollable by current traditional data-handling tools”. Contrary to the conventional data handling tools, the Big Data analysts at Apache and the research community developed a very efficient framework—called Hadoop—that can process and manage huge volumes of data [12]. Primarily, the Hadoop (Highly Archived Object-Oriented Programming) framework spawns the input data to multiple distributed computing nodes and provides reliable and scalable computing results [12, 13]. Bangari et al. [14] uses simple programming models based on Java language for distributed processing of a large volume of datasets through the clusters of computers. The basic idea of Hadoop setup is to use a single server in order to handle a collection of slave workstation nodes in which each node contains its own local storage and computational resources. To process and store data, Hadoop utilizes the MapReduce algorithm, which divides any given task into smaller parts in order to distribute them across a set of cluster nodes [15, 16]. Sharmila et al. [17] showed another basic feature of Hadoop as is its file system known as the Hadoop Distributed File System (HDFS), which is an efficient storage system for cost-effective hardware. Although Apache Hadoop remained one of the most reliable frameworks to handle Big Data within a decade after its first release in 2006, its efficacy was reduced after the exponential growth of streaming real-time data, Machine/Deep Learning (M/DL) technologies, and the use of graphics in online games & other related applications [18]. Bell et al. [19] overcome these limitations by another open-source framework called Apache Spark that was developed, which incorporates diverse features such as better memory & storage management and more scalability.

The primary objective of the research were analyzing the processing times and file management systems of the heterogeneous environment to get better performance of Hadoop. This paper contributes by giving a comprehensive comparison between Apache Hadoop & Apache Spark in terms of their application, scalability, reliability, security, cost-effectiveness, and other features. Moreover, the discussion of primary components of these frameworks ascertains their performance. The study concludes that the Spark framework is more useful for streaming applications, and hence it is fast and scalable. On the other hand, Hadoop has better security features, and it can handle very large volumes of data. Furthermore in this paper, a case study discusses the performance of the Hadoop framework by implementing the famous WorldCount algorithm. The results show the effectiveness of Hadoop in terms of processing time and storage required to process large dataset files. In addition, an analysis was described by comparing these results with Spark's framework to determine how its features could reduce the processing time of the algorithm for the given files.

The remainder of this paper is organized into the following sections. In Section 2, the characteristics of Big Data are discussed in detail. Section 3 provides the related work. Section 4 discusses the major components of Hadoop. Section 5 describes the WordCount Algorithm. Section 6 details the Spark framework followed by Spark Components in Section 7 and its comparison with Hadoop in Section 8. Section 9 describes the experimental setup and results by implementing the WordCount application on the Hadoop cluster. Then Section 10 discusses how Spark implementation of the WordCount application could improve the execution times. Finally, Section 11 provides the conclusion and future work.

II. CHARACTERISTICS OF BIG DATA

The Data grows in three dimensions—also known as the 3Vs model—according to the Gartner research report; these three are Volume, Variety, and Velocity [10, 11]. It has been observed that many industries and organizations use the 3Vs model to analyze Big Data. However, it cannot be formally defined by merely 3Vs, and many other characteristics also exist to properly define Big Data [20]. Chen et al. [10] stated these characteristics are extended to 5Vs including the above-mentioned 3Vs. These are elaborated on the basis of line spacing, and typestyles. Examples of these type styles are given below and are depicted in Figure 1.

A. Veracity

It is one of the most important properties of Big Data tools and is defined as data accuracy and its quality relative to its users. The veracity is ensured by providing accurate and clean data [10].

B. Volume

It may be defined as the bulk of data to be organized, stored, and processed. The data volume is exponentially growing, and it is expected to grow multiple times in the coming few decades. Chen et al. [10] Stated the current volume of Big Data generated per day is in the realms of Exabyte (1018).



Fig. 1. The 5Vs of Big Data.

C. Variety

It refers to the various forms in which data is available to the users. Currently, the data can be structured, unstructured, or semi-structured[11], depending upon its organization. Moreover, it may be in the form of plain text, image, audio, video, etc., or any combination of these forms.

D. Value

This simply implies how important or critical the data is to the user. Data value describes its beneficence for a particular organization or individual. The ratio of the valuable data is inversely proportional to the total volume of data. Chen et al. and Ward et al. [10, 11] for instance, in an hour-long video, the ratio of valuable data can be of a few seconds.

E. Velocity

Velocity implies the rate at which the data is retrieved, and it assists to identify the difference between normal data and Big Data. The characteristic of velocity for the data warehouse is a very significant parameter in this competitive field. For instance, it is the velocity that plays a critical role in data retrieval or storage at a typical social media warehouse for its users to efficiently use it for socializing. The author in [10, 11, and 17] found that the users of Facebook, Twitter, Instagram, and other famous platforms expect to communicate with each other in real-time for their everyday experience.

Most of the data analysts and experts suggest utilizing a variety of open-source Big Data platforms in order to take benefit from Big Data analytics [10, 21]. Elgendy and Elragal [22] showed these platforms offer a mixture of hardware resources using state-of-the-art software tools for data storage, processing, analysis, and visualization. The opportunities generally vary based upon the phenomenon where data is being utilized, and its value depends upon the type of applications. For example, in the stock exchange system, where demands and consumption change at a fast rate, data has importance only for a limited period of time [20]. Furthermore, due to the enormity of data volumes and Big Data applications, it becomes extremely tough for managers to select a single or a small group of data platforms. Undoubtedly, the ever-growing competition and limited time make it inconvenient for them to work with a large number of Big Data platforms. However, they still require multiple platforms due to the demands of

repeated and multiple task solvers [22]. Nonetheless, it can be a very useful analysis to determine an optimum set of platforms for a given organization considering its R&D requirements and applications.

Another significant aspect of the Big Data paradigm is data security concerns created during the management, storage, and processing of data.

F. Management Issues

Data is normally retrieved or stored in structured, unstructured, or semi-structured modes in various organizations. It is difficult to manage such diversity in data in large volumes.

G. Storage Issues

The data sources are diverse, and it is normally retrieved from social media, streaming systems, mobile signal coordinates, sensor information, online recordings, and e-business exchange reports. Storing this data in various forms creates storage issues and requires standards to be properly implemented.

H. Processing Issues

Based on user requirements, the Big Data systems are expected to process data in volumes of Petabyte, Exabyte, or even in Zettabyte. This processing can be real-time or in batch mode. Therefore, Big Data systems must be able to cope with user requirements.

I. Security Issues

In government & private sectors, the data is normally vulnerable to malicious attacks and intrusions. Therefore, organizations are expected to carry effective intrusion detection systems and data integrity systems in order to ensure the safety of user data and avoid data exploitation.

All the above-mentioned issues and challenges are tackled by using efficient tools like Apache Hadoop & Spark. Currently, the most widely used framework is Hadoop, and it is particularly useful for processing large volumes of data by tech giants such as Twitter, LinkedIn, eBay, and Amazon [23]. A lot of research is performed to evaluate the performance of Hadoop, but there is a need to improve its functionality in terms of its time efficiency. [16,18] Identified that Hadoop is extremely powerful in the case of storage systems due to HDFS, but it struggles to compete with Spark in the processing part performed by its MapReduce algorithm. This work focused on the testing of MapReduce by recording the time elapsed by each processing step of the algorithm with various volumes of data downloaded in the form of data files from the Internet. The primary aim is to ascertain the performance of Hadoop compared to Spark to learn which application scenarios are suitable for a particular framework.

III. RELATED WORK

Due to the enormous amount of data generated daily, the issues of its management, storage, and processing are not only significant for the academic community but also the industry. For instance, Zhao et al. [18] conducted a performance comparison between Hadoop and HAMR based on the running PageRank algorithm. HAMR is a new technology, which

provides faster processing and memory utilization compared to Hadoop. The comparison parameters used in the research were memory usage, CPU consumption, and running time. Shah et al. [24] observed the performance of Hadoop in a heterogeneous environment with various types of hardware resources. They developed an algorithm called Saksham, which enables the rearrangement of the data blocks to optimize the performance of the Hadoop in homogeneous & heterogeneous environments. A region-based data placement policy is proposed by Muthukkaruppan et al. [25]. The main purpose of the proposed policy is to achieve high fault tolerance and data locality, which does not exist in the default policy. A specific region data block is placed in the contiguous data portion of nodes in the region-based policy. Qureshi et al. [13] described storage media-aware policy in order to improve the performance of Hadoop. This policy is known as the Robust Data Placement (RDP), which also handles the network traffic and unbalanced workload.

Meng et al. [26] proposed a strategy that places data blocks with disk utilization and network load while in default Hadoop, block placement is done by Round Robin that reduces the performance in a heterogeneous environment. This strategy improved the HDFS performance by enhancing the space storage utilization and throughput.

Similarly, Dai et al. [27] presented their proposed Replica Replacement Policy (RRP) developed in 2017 to improve the Hadoop performance by eliminating the utility of HDFS balancer; consequently, the replica is evenly distributed among the homogeneous and heterogeneous nodes. This policy generates better results of replica management as compared to the default replica management policy of the HDFS in Hadoop. Herodotou et al. [28] proposed a new tool to optimize the default parameters of Hadoop; for instance, the total number of map reduces, scheduling policy and the reuse of JVM to increase the performance. The tool is called Startfish, and its main purpose is to work with Hadoop phases such as placement, scheduling, and tuning of the assigned jobs to the computing nodes. Panatula et al. [29] worked on the performance of HadoopMapReduce Word Count Algorithm and presented it on the Twitter data. The experimental setup was based on a 4-node Hadoop system to analyze the performance of the algorithm. It was concluded that Hadoop can work efficiently with the setup of 3 or more nodes. Gohil et al. [15] processed a different set of applications of MapReduce including Word Count and Tera Sort etc. to evaluate Hadoop performance. The evaluation parameters were to set up a dedicated cluster and decrease the I/O latency of the network. Likewise, the in-house Hadoop cluster setup and Amazon EC2 instances are also used to evaluate the Hadoop performance. Khan et al. [30] have modeled the estimation of the provisioning of the resources and completion time of the jobs. Furthermore, the Hadoop and Spark-based distributed system performance has been evaluated by Taran et al. [31]. A performance evaluation framework is proposed for Hadoop by Lin et al. [16] on the basis of cluster computing nodes across a clustered network. A configuration strategy is proposed by Jain et al. [32], in which MapReduce parameters are configured with optimized tuning to improve the performance of Hadoop. To analyze the performance of Hadoop, several applications

were processed and tested repeatedly by Londhe et al. [33] using the Amazon platform for Hadoop. In [34], an analysis of the computational performance of the processors on a private network was presented in order to reduce the input/output latency of the network.

This work implemented a well-known WordCount Algorithm using Hadoop to evaluate the framework performance and provide a thorough analysis of the difference between Hadoop and Spark frameworks.

IV. HADOOP COMPONENTS

Hadoop framework allows the users to record & process Big Data in a distributed network, across several computing nodes using easy-to-use programming methods. It is an open-source framework designed by D. Cutting & M. Cafarella [14]. Most researchers and developers consider Hadoop the most efficient tool in the Big Data domain. It is sometimes misunderstood as merely a database, but it is a comprehensive ecosystem that allows distributing data for processing across thousands of servers and keeping the overall performance extremely optimized. As mentioned earlier, there are two basic components of the Hadoop system: HDFS and MapReduce algorithm [35].

The basic architecture of the Hadoop framework is depicted in Figure 2. It based on the Master-Slave system, in which the master node is called the name node, and the slave is the data node, which keeps and processes the actual data. For fault tolerance, the factor of replication is set at 3, where the MapReduce algorithm helps the replicated data to be processed in parallel mode [26]. In the following, the components of the Hadoop system describe in detail:

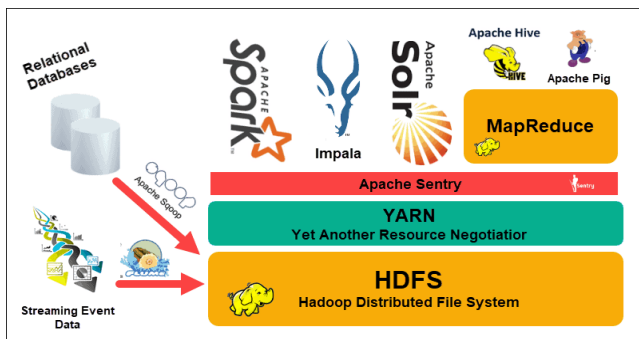


Fig. 2. Hadoop Architecture.

A. HDFS

A large amount of data in the form of sets is stored on HDFS which is a distributed file management system and works on the commodity hardware [38]. Thousands of nodes clustered in a distributed system can be supported using the HDFS in a reliable manner at a low cost. It can support large files containing volumes of data in terabytes. Furthermore, it provides portability for data across various platforms and nodes. However, the most important feature of HDFS is its ability to reduce traffic congestion across networks, because processing and data are moved closer to each other.

To perform low-latency computations, another significant open-source system has been developed on top of HDFS,

which is called the HBase [36]. It is essentially a distributed non-relational database system using column-based value/key data presentation. [36, 37] found a developer can spawn data across distributed cluster nodes and update the data tables at very fast rates by using HBase. However, HDFS cannot be replaced and mixed with HBase due to its non-relational DBMS nature, although HBase can help to process the real-time data by using its in-memory processing engine [40]. The HDFS is used in many systems that consist of conventional file systems such as ext2 (Linux) or FAT (Windows). But HDFS is totally different from all these traditional file systems for the following main reasons:

HDFS is optimized to maximize data rates. The size of a data block is 64MB in HDFS as compared to 512 bytes to 4 KB in most of the traditional file systems, which significantly reduces the seek time. In addition, it is possible to further extend the size of a block to 128MB or 256MB. Moreover, [26] showed HDFS is a Write Once Read Many (WORM) file management system: any file can be written once but can be accessed several times.

Another prominent feature of HDFS is its fault tolerance where it furnishes a block-based replication framework with a configurable number of replications (by default, it is 3) [23, 26]. During the composing stage, every block compared to the record is imitated on isolated hubs in the bunch, which assists with ensuring its dependability and accessibility, when understanding the information. In the event that a block is inaccessible on one node, duplicates of that block are ensured to be accessible on a different node.

HDFS builds on the native OS file system to present a unified storage system built on a heterogeneous array of disks and file systems.

B. MapReduce

White [38] found MapReduce is the most significant component of the Hadoop framework. In the clusters of the Hadoop, the scalability is provided by MapReduce on thousands of computing servers. The term 'MapReduce' is constituted from two different words; Map and Reduce. Both these terms are attributed to performing different tasks in the Hadoop system. The job of mapping data is performed by the Map function. The map function converts the original data into the form of sets. The purpose of making sets is to make key pairs of all unique elements of data, which serves as the output of the map function. Subsequently, the output of the Map function becomes the input data of the Reduce function. The main job of the Reduce function is to reduce the number of key pairs of all unique elements in a key pair [15]. Figure 3 depicts the MapReduce process. The input data is split by the distributed file system and mapped to the Map node in key/value format. Then the Reduce node merges the key/value pairs to further reduce them and store them using the file system [38].

1) *In the Hadoop system*, the job of the Map function is performed firstly, followed by the Reduce function. Sometimes the map function job is enough to process the data efficiently. In Figure 3, the architecture and process of the MapReduce give a thorough overview.

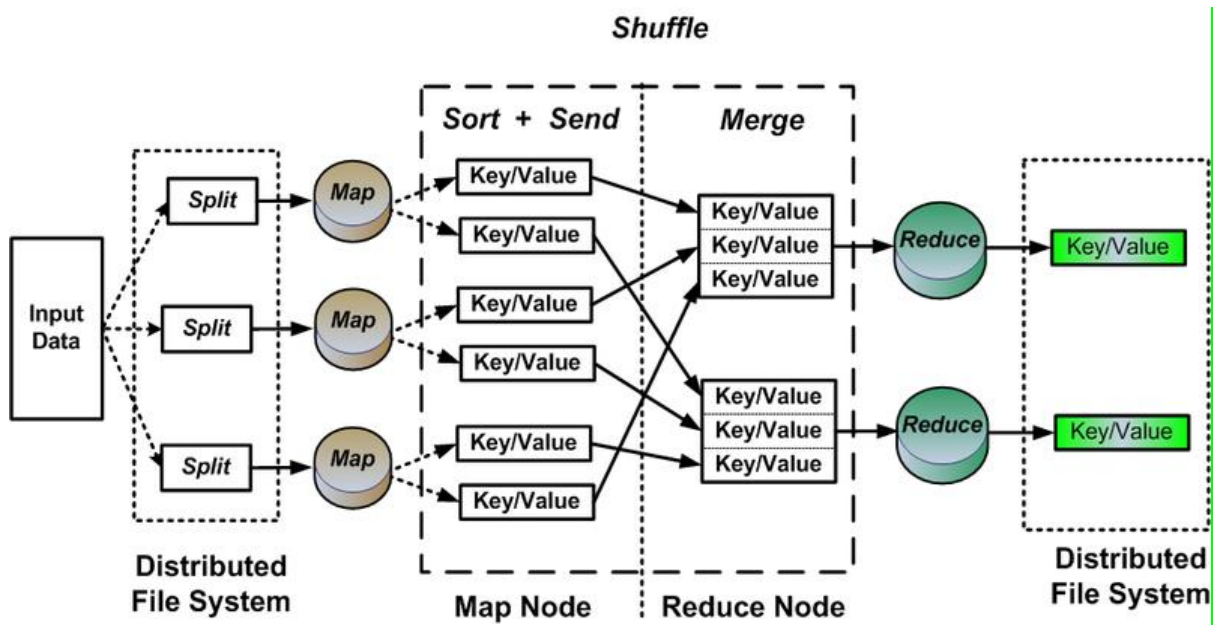


Fig. 3. MapReduce Process [39].

2) The tasks are distributed on multiple nodes in the MapReduce, which works as the computational model of Hadoop in order to process a large amount of data on clusters containing a large number of servers/nodes.

3) MapReduce works as the processing component of Hadoop because it processes the data concurrently on multiple nodes to reduce the overall computational cost and henceforth, increase the efficiency. When MapReduce gets data as input, it distributes it on the different nodes. The output of the mapper becomes intermediate data, which is in the split form. Subsequently, the shuffle process is used to exchange the data between various nodes. The data containing the same key is assigned to the same key of the reducer node and the output of the reducer is finally stored.

C. Hadoop WordCount

Lin, and Liu; Sharmila *et al.* [16, 17] stated that the Hadoop WordCount is a simple algorithm that is used to read the input text files, and count the number of unique words existing in a file.

Application	System Resource Utilization
WordSort	Sort Phase: IO-bound, Reduce Phase: Communication-bound.
Word Count	CPU-bound
TeraSort	Map Stage: CPU-Bound Reduce stage: IO-bound
NutchIndexing	IO-bound with high CPU utilizations in map stage. This workload is mainly used for web searching.
Kmeans	CPU-bound in the iteration phase, and IO-bound in the clustering phase. It is used for machine learning and data mining.

Fig. 4. Hadoop WordCount Resources Utilization [15].

In this algorithm, the input is a text file and different output files are generated in which information about words and their count is computed. On each level of this algorithm, different computing resources are utilized as depicted in Figure 5. In the Hadoop WordCount, the map and reduce functions perform in parallel. Figure 4 shows a comparison between some of the famous applications which are used to test the Hadoop system. It also presents the system resource utilization in the case of each application, and it can be noted that Word Count is a CPU-bound application.

V. THE WORD COUNT ALGORITHM

Figure 5 depicts WordCount Algorithm for counting the number of words occurrences in a file by using MapReduce programming.

```

1: Class Mapper <K1,V1,K2,V2 >
2: function map(K1,V1)
3: List words = V1.splitBy (token) ;
4: foreach K2 in words
5: write(K2,1) ;
6: end for
7: end function
8: end class

1: Class Reducer <K2,V2,K3,V3 >
2: function reduce(K2,Iterable<V2> itor)
3: sum = 0 ;
4: foreach count in itor
5: sum += count ;
6: end for
7: write(K3,sum) ;
8: end function
9: end class

```

Fig. 5. Algorithms of Word Count in Hadoop.

The principle is the same as above to count the integers: build pairs (key, list) where "key" is a word and "list" a list of 1 each designating an occurrence of a word.

The map step includes additional work which consists of breaking up the text file (or a simple character string) into words. A word is simply located between two patterns or the 'space' characters. The combined step is implicit before executing the reduce function toolbar. In Figure 5, the main job of the map function is to split the words of a line at a time through the tokens. A key value is assigned to each unique word and the output of this function is in the form of a key/value set, which contains the word and a key-value with the format: "<<word>, 1>". These key/value pairs are the inputs of the Reduce function. In the default order of the dictionary, the keys are sorted before the execution of the Reduce function. The reduce function job is to count the occurrence of each unique word. Finally, the Reduce function outputs the result on HDF. An example of the Map and Reduce function is shown in Figure 6. The original data is stored in a file containing different words, and after the WordCount algorithm is applied, the output is shown in the form of word counts.

In this example (depicted in Figure 6), the original data is shown in the left block. This data is retrieved from a file and mapped on the Map function. This file has different words such as Test, Reduce, Map, and HDFS. The map function splits these words using tokenization. In the map function, a key/value set is assigned to each unique word of the file which creates the key/value pairs and serves as the input of the Reduce function. The counting process is done by the Reduce function. The Reduce function count of each unique word is taken as the output, and it is stored on HDF to be later saved on the local storage.

Similar to the WordCount, there are many other applications of MapReduce, such as TeraSort and Sort. In the Hadoop MapReduce, TeraSort is used to sort the 1TB of data at an extremely fast rate. The HDFS file system makes it an ideal choice to fine-tune the configuration of a Hadoop cluster to quickly process these applications. In MapReduce, sort is an algorithm with an objective to process and analyze the data. According to [39], the improvement in MapReduce application is achieved at 59.15813, 64.23517, and 18.05475 for the TeraSort, Sort, and WordCount applications respectively, if the default settings of Hadoop at map level are considered.

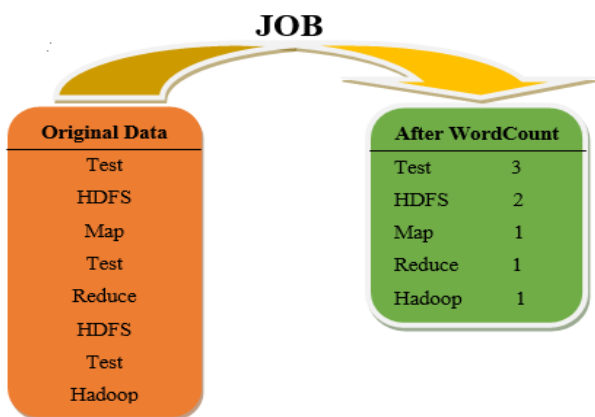


Fig. 6. Data Sample of WordCount.

VI. SPARK BASED BIG DATA ANALYSIS

Bell [19] establishes that the Apache Spark is a model-based open-source framework that is used to examine huge datasets in streaming applications. This framework was established in contrast with the Hadoop MapReduce model at UC Berkeley AMPLAB [19, 40]. Architecture of Apache Spark is shown in Figure 7. Bell [19] showed that the main components of Apache Spark are the program of driver, initiators, cluster director, and the HDFS. The main program of Spark is its driver program. At the time of the startup of the Spark program, Spark Context is produced that plays a significant role in the whole implementation of the job [38]. The resources are managed in clusters when the Spark Context program is linked with the cluster manager. In order to store the app information and run the logic, the program is spawned through the cluster managers.

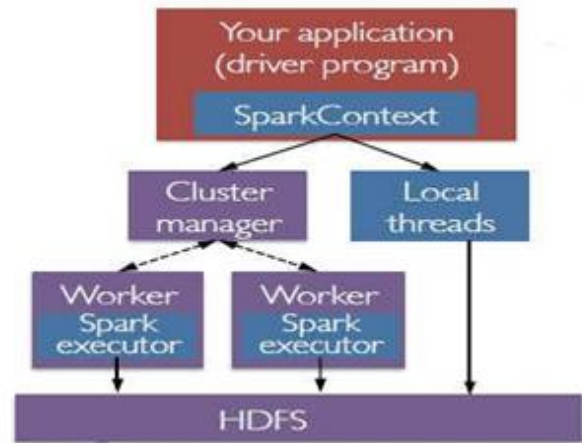


Fig. 7. Architecture of Spark.

Bell [19] said that Spark can be operated using multiple programming languages such as Scala, Java, Python, and R. Each language is provided with multiple libraries & APIs to support Big Data analytics. However, the most popular language in the Spark community is Scala, as Spark's core is implemented in Scala. Moreover, scalability is always a great concern in the case of Big Data analytics, which is handled efficiently in the Scala language. It is more compact compared to Java; a program written in Java is much larger in size compared to its equivalent of Scala. For comparison, it can be noted that one line of code in Scala is equal to 20-25 lines of code in Java. In order to enhance efficiency and reliability, numerous programs are being transformed from Java to Scala [33].

Bell [19] stated that the basic framework of Spark consists of two major technologies namely, Resilient Distributed Dataset (RDD) & Directed Acyclic Graph (DAG). The description of these two technologies as follows:

A. Resilient Distributed Datasets (RDD)

The Resilient Distributed Datasets (RDDs) are used to gather elements that are fault-tolerant and continue in parallel fashion as a primary concept of Spark [41]. Once the RDDs are created, they cannot be changed. It is impossible to change

them even with their ability to transform and perform actions. These datasets help to reorganize the computations and data processing enhancement [41]. Ganesh *et al.* [29] found a typical RDD provides information to regenerate on multiple nodes, and that is the primary reason for their fault tolerance. By transforming the current lists of data or changing the files in HDFS, the RDD is created. In case of misinformation of specific compartments of RDD, the default value is used by a spark programmer. A typical RDD executes two types of methods given as follows:

B. Transformation

A new RDD instead of a single value run at the time of performing changes on RDD. The analyses of computation are not sudden to transform, because the computation speed is very slow. When a program runs on this, then they are implemented. The following functions of transformation are performed: Mapping, Filtering, Reduce ByKey, FlatMap, and Group ByKey.

1) *Action*: A single value is examined and run when action methods are used on RDDs. At the time of an action method, the computation of data processing is performed and sent to a resultant value. In this respect, few action techniques are primary, take, decrease, gather, count, for each, and Count ByKey.

C. Directed Acyclic Graph (DAG)

Directed Acyclic Graph (DAG) engine is an updated Spark method to help cyclic data flow [40,41]. DAG consists of task phases that should be performed on the cluster of Spark nodes. Spark produces several DAGs of the input data that consist of an arbitrary number of steps, while DAG is further processed by the MapReduce which consists of two steps of Map and Reduce, as described in Section IV. Gohil *et al.* [15] showed that after the completion of a single step, this process allows for the processing of a simple task as compared to a complex task to be processed in multiple steps in a single run.

VII. COMPONENTS OF APACHE SPARK

Spark's underlying architecture is just like Hadoop, but it uses the in-memory system to process the streaming and graphics data. Due to the provision of in-memory processing, it can handle complex analyses on large volumes of data. There are multiple subsystems for memory handling, job assignment, fault tolerance, and storage, etc. Moreover, Spark can access information stored on different platforms such as Hadoop HDFS, Mesos, Mongo DB, Cassandra, H-Base, Amazon S3, and the data sources from standard cloud interfaces [34]. Figure 8 depicts major components of the Spark system. These components include Spark SQL, Spark Streaming, MLlib, and GraphX [16]. The brief discussion these components are as follows:

A. Directed Acyclic Graph (DAG)

The primary component of this framework is Spark core. It handles all the basic functions related to I/O, dispatching, and arrangement of the distributed tasks. All the other components interact with the core to invoke their basic functionalities.

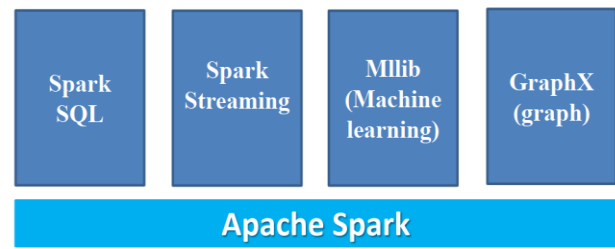


Fig. 8. Components of Spark.

B. Spark SQL

Spark SQL combines the traditional relational tables with RDDs for users to use SQL commands to communicate with large datasets for complex analytics [42]. It permits the use of old BI and visualization tools to allow performing SQL queries on Spark. An advanced RDD data concept method is introduced to provide support for structured and semi-structured data.

C. Spark Streaming

Kroß and Krcmar [43] stated that Spark Streaming is used for the processing of real-time data. To support the processing of real-time data of RDD, it uses an API called DStream. This component has the ability to seamlessly parallelize the data for streaming applications.

D. Spark GraphX

GraphX is a library that introduces 'the Resilient Distributed Property Graph' introduced by GraphX which has the ability to work with graphics and related computations [44]. Aggregate messages, sub-graphs and join vertices, and an optimized variant of Pregel API are various operators of GraphX. It also provides various graph-related algorithms and builders to simplify the graph analytics tasks.

E. MLlib

Meng *et al.* [45] found MLlib is an extensive framework to invoke Machine Learning (ML) options in Spark. It enables several basic ML algorithms like clustering, classification, and regression, etc. In Big Data, mining requires automation-based systems to dig out certain patterns of data, and MLlib provides a platform for data mining.

MLlib is a relatively new technology in Spark as compared to Mahout [46], which has been utilized as a set of Java APIs for machine learning. Mahout provides several optimized ML algorithms, thereby relieving its users to worry about the development of basic algorithms. Instead, they can work on their analytical problem on their datasets. The basic advantage of MLlib over Mahout is that it offers regression models [45,46].

VIII. HADOOP VS SPARK

In the Big Data domain, the need for efficient, accurate, and reliable analytics along with excellent data visualization is increasing with every passing day [10, 11, 17]. Geczy [47] Identified that Tech giants like Yahoo, Amazon, Uber, eBay, Facebook, and Twitter, etc. heavily rely on these analytics for their smooth operations. These companies need optimized costs, real-time analytics, and fault tolerance with their data

processing. For more than a decade, Hadoop served as a prominent platform in order to provide these options to tech companies, government organizations like NASA & CERN, but the technological needs are becoming more and more advanced. Nowadays, Big Data analytics also require Machine Learning (ML) capabilities, streaming systems, and graphics processing [43–45]. Hadoop is an excellent platform for batch processing and large volumes of data, but it struggles with streaming applications and Machine Learning capabilities, etc. [10, 45]. Ward, and Barker [11] showed to overcome these limitations in Hadoop, Spark was introduced as another open-source project. It has built-in libraries for streaming, graphics, and ML capabilities, and it is designed to work on multiple programming languages.

Currently, both Hadoop and Spark are being used globally for different reasons, and they are the most prominent platforms for Big Data analytics and processing. Therefore, it is very hard to figure out which of these two is better. In [35], a comparison in terms of optimization is presented, and it is concluded that Spark performs much better in terms of accessing the storage systems and utilization of memory bandwidth. Table 1 present a brief comparison of Hadoop with Spark in terms of various features. The selection of these features from state-of-the-art research [2].

As shown in Table 1, the major difference between Hadoop and spark is the underlying methodology used to process the data [48]. Spark has the capability to perform Batch & Stream processing, whereas Hadoop can only operate with Batch processing. This particular capability is invoked in Spark by allowing in-memory processing, instead of disk-based computations in Hadoop. Hence, in the case of Hadoop, read and write operations are performed by seeking data available on a disk. This provides a fundamental advantage to Spark over Hadoop, as this difference in the technique of processing greatly affects the processing speeds. As all the processing in the case of Spark is performed in the memory, it takes lesser time to process data. According to [35], Spark is 100 times faster as compared to Hadoop.

Hadoop is most suitable in applications where the results are not immediately required in a short time. Contrarily, Spark is useful in applications when the results are needed in real-time. For instance, Spark is a good platform for stock price evaluation and enterprises [35].

Another major difference between the two frameworks is the amount of data to be processed. Htay and Phyu [39] indicated that Hadoop can process a massive amount of data as compared to Spark. Initially, Hadoop was also utilized for data archiving because Hadoop has the ability to store the data in large volumes due to HDFS. Using Hadoop, a user can store the data not only for a few years but for decades in its original and archiving form [1]. Chen and Zhang [7] stated that spark is simple to use as compared to Hadoop due to its easy-to-use APIs and many friendly features. If cost is taken in consider, Hadoop is less costly as compared to Spark. Similarly, Hadoop has better security features compared to Spark. Hadoop does not support data caching whereas Spark supports caching of data memory. Hadoop has a higher latency of computing as it has less data interactivity, as opposed to Spark [48].

As presented in Table 1, both Hadoop and Spark support auto-scaling, which means that if the data requires more nodes, then the frameworks give provision to add more nodes for distributing the workload automatically. Furthermore, as discussed above, Spark is flexible to support multiple programming languages in contrast to Hadoop only allowing Java programming.

Based on this comparison, it is very difficult to determine which framework is better. But generally, it can ascertain that both are useful for their own set of applications. For instance, Spark is better in real-time applications, where streaming data is required to be processed, whereas Hadoop is useful for processing large volumes of data when it is not strictly time-bound [1]. Furthermore, Spark provides excellent built-in APIs and libraries to optimize most of the user tasks, which can help to save a lot of programming effort. It has higher interactivity of data. Finally, it offers options for batch processing and streaming, as well as ML & graphic processing. Hence, Spark is obviously more advanced.

TABLE I. COMPARISON BETWEEN SPARK AND HADOOP

Features	Hadoop	Spark
Processing Mode	Batch	Batch and Stream
Scalability	Horizontal	Horizontal
Message Delivery Guarantees	Exactly once	Exactly once
Computation Mode	Disk-based	In memory
Auto-scaling	Yes	Yes
Iterative Computation	Yes	Yes
Speed	Slow	Fast
Amount of Data for Processing	More	Less
Security	More Secure	Less Secure
Cost	High	Low
Performance	Fair	Good
Language	Scala	Java
Data Caching	Support	No Support

IX. EXPERIMENTS AND RESULTS

To evaluate the performance of the MapReduce algorithm, the study used the Hadoop framework by implementing the WordCount algorithm in Java and performed the corresponding data analysis. The dataset consists of four different text files named chas.txt, ac.txt, xad.txt, and xaa.txt; all downloaded from the Internet. The corresponding file sizes are 202MB, 137MB, 69MB, and 34MB respectively. This test dataset contains random text comprising diverse words. Table 1 presents the study dataset.

To execute the data, a number of computing machines (or nodes) clustered across a network work together in MapReduce fashion. This number can be as large as in millions if the volumes of data are extremely large. The Map and Reduce functions may work on the same machine or on separate machines. The management of these machines is performed by master node(s).

A developed of a cluster of three machines to crunch the data given in Table 2 that used in the experiment. The hardware resources of the setup are presented in Table 3. The master node had four cores with main memory of 2GB, and it handled all the control processes. The other two nodes were single-core machines with 1GB of memory each.

The cluster executed the WordCount application, and it recorded the total time elapsed by the MapReduce function with respect to the size of each file. The algorithm first loads the text data into the HDFS data storage. When this data is available on the cluster, it can be used for analysis. Subsequently, the algorithm reads the text files and counts all the words in the given text files using MapReduce functions. The results are presented in Table 4. The investigation evaluated the time of Map and Reduce functions separately in the case of each file. The file sizes of the 202MB, 137MB, 69MB, and 34MB take 954645ms, 153826ms, 86224ms, and 57508ms to map all the words in each file by the Map function, respectively. On the other hand, the Reduce function takes 57964ms, 51089ms, 5095ms, and 6472ms, respectively.

TABLE II. TEST DATASET FOR WORD COUNT ALGORITHM

File name	File size (in MB)
chast.txt	202
ac.txt	137
xad.txt	69
xaa.txt	34

TABLE III. HARDWARE RESOURCES FOR TEST CLUSTER

Machine	Cores	Memory	Network
Master	4	2GB	192.168.1.4
Slave1	2	1GB	192.168.1.6
Slave2	2	1GB	192.168.1.7

TABLE IV. HADOOP-MAPREDUCE (WORDCOUNT) IMPACT ON FILES SIZE

Data File (in MB)	Time elapsed by Map function (ms)	Time elapsed by Reduce function (ms)
202 (chast)	954645	57964
137 (xac)	153826	5108
69 (xad)	86224	5095
34 (xaa)	57508	6472

Figure 9 shows the graphical representation of the MapReduce function time on the WordCount algorithm. The slope of the graph shows that the map function takes more time compared to the reduced function for the WordCount algorithm. It can be noted that the time elapsed by the Map and Reduce functions are directly proportional to the size of the file. As the size of the file increases, the time elapse of the Map and Reduce functions is also increased. The slope of the reducer function goes upward slightly as the size of the file increases, but the time slope of the mapper function goes upward at 75 degrees as the file size increases from 137MB to 202MB. The difference between these two files is not too

much. The performance of the mapper function is not very prominent in case of an increase in the file size. This is because Hadoop is normally optimized for the data file sizes in GBs and TBs. Another reason for the timing results is the size of the cluster that have used. Hadoop is normally utilized for clusters with thousands of nodes, crunching data volumes in the range of gigabytes & terabytes, etc.

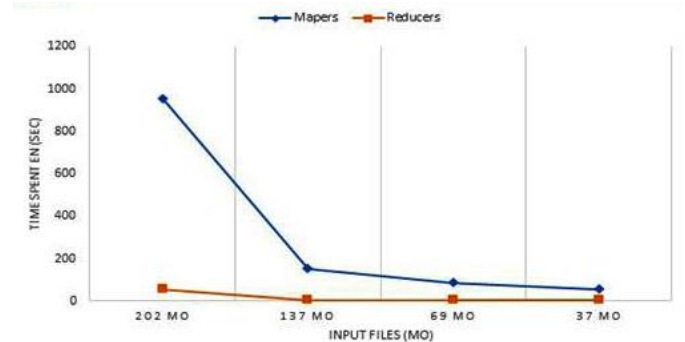


Fig. 9. WordCount Algorithm Time Impact.

X. DISCUSSION

The study implemented the WordCount application using Hadoop and executed it in a cluster, as discussed in Section IX. If all the parameters and hardware resources are kept the same, and the WordCount algorithm is implemented on the Spark framework, the study note that the overall processing time is reduced by a factor of 10. Similarly, the in-line memory processing option in Spark makes it 100x faster for memory operations. Another advantage of Spark is the reduced number of Lines of Codes (LoCs) of Scala. The Hadoop implementation took 60 LoCs in Java to write the MapReduce function, but the same code took only 5 LoCs in Scala in the case of Spark. Therefore, the only complication in the case of Spark is learning a Scala, but this learning can reduce a lot of effort.

XI. CONCLUSION AND FUTURE WORK

This article described two frameworks, Hadoop and Spark, which allow the processing of Big Data on clusters of computing machines. The study first discussed the main characteristics of Big Data followed by the primary components of Hadoop and Spark. Then it discussed the underlying storage architecture called HDFS in both frameworks. Afterwards it detailed the MapReduce algorithm used as a core program in Hadoop operations. Subsequently, it differentiated both the frameworks on the basis of various features like cost, scalability, programming languages, and processing modes, etc. The enquiry concluded that Spark is better for streaming applications, while Hadoop is better in the case of processing large datasets in batch processing mode. One of the main objectives of this work is to evaluate the performance of the WordCount application in terms of processing time for both frameworks. Finally, the study provided a discussion on the limitations of Hadoop as compared to Spark while processing the WordCount application. The investigation concludes that the performance of Hadoop can be measured on the basis of different aspects, such as tuning of the MapReduce parameters and the total

number of nodes, etc. In the future, it intend to extend this work by taking big datasets on a larger cluster and report the memory speed and execution times for more complex applications. The research will compare the performance of Hadoop and Spark for a given set of application parameters, and propose an optimization function to choose a particular framework for a given application. It will evaluate the performance of these applications in terms of data growth, the number of iterations, and data processing in real-time. Unfortunately, the current version of Spark is unable to handle a large number of workloads using SQL-like queries, as required by tech giants like Google, Facebook, and Yahoo, etc. Consequently, some modifications are suggested in newer Spark versions to allow multithreading options to spawn application tasks more efficiently. The research intends to further explore Spark to find out optimized options for application submission using large datasets.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

- [1] Z. Tufekci, “Big Questions for social media big data: Representativeness, validity and other methodological pitfalls,” Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, pp. 505–514, 2014.
- [2] R. Sakthivel, V. Parthipan, and D. Dhanasekaran, “Big data analytics on smart and connected communities using Internet of Things,” International Journal of Pharmacy and Technology, vol. 8, no. 4, pp. 19590–19601, 2016.
- [3] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, “Deep learning for IoT big data and streaming analytics: A survey,” ArXiv, vol. 20, no. 4, pp. 2923–2960, 2017.
- [4] U. Srivastava, and S. Gopalkrishnan, “Impact of big data analytics on banking sector: Learning for Indian Banks,” Procedia Computer Science, vol. 50, pp. 643–652, 2015, doi:10.1016/j.procs.2015.04.098.
- [5] <https://www.socialmediatoday.com/news/10-social-media-statistics-you-need-to-know-in-2019-infographic/559181/>.
- [6] <https://www.statista.com/markets/>.
- [7] C.L. Philip Chen, and C.Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data,” Information Sciences, vol. 275, pp. 314–347, 2014, doi:10.1016/j.ins.2014.01.015.
- [8] S. Venkatraman, K.F.S. Kaspi, and R. Venkatraman, “SQL Versus NoSQL Movement with Big Data Analytics,” International Journal of Information Technology and Computer Science, vol. 8, no. 12, pp. 59–66, 2016, doi:10.5815/ijitcs.2016.12.07.
- [9] T.K. Das, and P. Mohan Kumar, “Big data analytics: A framework for unstructured data analysis,” International Journal of Engineering and Technology, vol. 5, no. 1, pp. 153–156, 2013.
- [10] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.
- [11] J.S. Ward, and A. Barker, “Undefined By Data: A Survey of Big Data Definitions,” ArXiv Preprint ArXiv:1309.5821, 2013.
- [12] D. Zhao, “Performance comparison between Hadoop and HAMR under laboratory environment,” Procedia Computer Science, vol. 111, pp. 223–229, 2017, doi:10.1016/j.procs.2017.06.057.
- [13] N.M.F. Qureshi, and D.R. Shin, “RDP: A storage-tier-aware robust data placement strategy for hadoop in a cloud-based heterogeneous environment,” KSII Transactions on Internet and Information Systems, vol. 10, no. 9, pp. 4063–4086, 2016, doi:10.3837/tiis.2016.09.003.
- [14] K. Bangari, S. Meduri, and C.C.Y. Rao, “Implementation of Word Count-Hadoop Framework with Map Reduce Algorithm,” International Journal of Computer Trends and Technology (IJCTT), vol. 49, no. 3, pp. 179–182, 2017.
- [15] P. Gohil, D. Garg, and B. Panchal, “A performance analysis of MapReduce applications on big data in cloud based Hadoop,” in 2014 International Conference on Information Communication and Embedded Systems, ICICES 2014, IEEE: pp. 1–6, 2015, doi:10.1109/ICICES.2014.7033791.
- [16] W. Lin, and J. Liu, “Performance analysis of MapReduce program in heterogeneous cloud computing,” Journal of Networks, vol. 8, no. 8, pp. 1734–1741, 2013, doi:10.4304/jnw.8.8.1734-1741.
- [17] K. Sharmila, S. Kamalakkannan, R. Devi, and C. Shanthy, “Big data analysis using apache hadoop and spark,” in International Journal of Recent Technology and Engineering, IEEE: pp. 167–170, 2019, doi:10.35940/ijrte.A2128.078219.
- [18] V. Kalavri, and V. Vlassov, “MapReduce: Limitations, optimizations and open issues,” in Proceedings - 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2013, IEEE: pp. 1031–1038, 2013, doi:10.1109/TrustCom.2013.126.
- [19] J. Bell, “Apache Spark,” Machine Learning, vol. 17, pp. 275–314, 2015, doi:10.1002/9781119183464.ch11.
- [20] B. Data, “Big data characteristics and sources,” The Macrotheme Review, vol. 3, no. 6, pp. 8–10, 2017.
- [21] Begum, F. Fatima, and R. Haneef, “Big Data and Advanced Analytics,” in Advances in Intelligent Systems and Computing, IEEE: pp. 594–601, 2019, doi:10.1007/978-3-030-11890-7_57.
- [22] N. Elgendy, and A. Elragal, “Big data analytics: A literature review paper,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer: pp. 214–227, 2014, doi:10.1007/978-3-319-08976-8_16.
- [23] J.R. Saura, B.R. Herraiez, and A. Reyes-Menendez, “Comparing a traditional approach for financial brand communication analysis with a big data analytics technique,” IEEE Access, vol. 7, pp. 37100–37108, 2019, doi:10.1109/ACCESS.2019.2905301.
- [24] Shah, and M. Padole, “Saksham: Resource Aware Block Rearrangement Algorithm for Load Balancing in Hadoop,” Procedia Computer Science, vol. 167, pp. 47–56, 2020, doi:10.1016/j.procs.2020.03.181.
- [25] K. Muthukkaruppan, K. Ranganathan, and L. Tang., U.S. Patent Application No. 14/996,627, 2016.
- [26] L. Meng, W. Zhao, H. Zhao, and Y. Ding, “A network load sensitive block placement strategy of HDFS,” KSII Transactions on Internet and Information Systems, vol. 9, no. 9, pp. 3539–3558, 2015, doi:10.3837/tiis.2015.09.014.
- [27] W. Dai, I. Ibrahim, and M. Bassiouni, “An Improved Replica Placement Policy for Hadoop Distributed File System Running on Cloud Platforms,” in Proceedings - 4th IEEE International Conference on Cyber Security and Cloud Computing, CSCloud 2017 and 3rd IEEE International Conference of Scalable and Smart Cloud, SSC 2017, IEEE: pp. 270–275, 2017, doi:10.1109/CSCloud.2017.65.
- [28] H. Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F.B. Cetin, and S. Babu, “Starfish: A Self-tuning System for Big Data Analytics,” in Cidr, pp. 261–272, 2011.
- [29] P. Ganesh, K. Sailaja Kumar, D. Evangelin Geetha, and T. V. Suresh Kumar, “Performance evaluation of cloud service with hadoop for twitter data,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 13, no. 1, pp. 392–404, 2019, doi:10.11591/ijeecs.v13.i1.pp392-404.M. Khan, Y. Jin, M. Li, Y. Xiang, and C. Jiang, “Hadoop Performance Modeling for Job Estimation and Resource Provisioning,” IEEE Transactions on Parallel and Distributed Systems, vol. 27, no. 2, pp. 441–454, 2016, doi:10.1109/TPDS.2015.2405552.
- [30] V. Taran, O. Alienin, S. Stirenko, Y. Gordienko, and A. Rojbi, “Performance evaluation of distributed computing environments with hadoop and spark frameworks,” in arXiv, IEEE: pp. 80–83, 2017.
- [31] Jain, and M. Choudhary, “Analyzing & optimizing hadoop performance,” in Proceedings of the 2017 International Conference On

- Big Data Analytics and Computational Intelligence, ICBDAI 2017, IEEE: pp. 116–121, 2017, doi:10.1109/ICBDACI.2017.8070820.
- [32] S. Londhe, and S. Mahajan, “Effective and Efficient Way of Reduce Dependency on Dataset With the Help of Mapreduce on Big Data,” *International Journal of Students’ Research in Technology & Management*, vol. 3, no. 6, pp. 401, 2015, doi:10.18510/ijstrm.2015.364.
- [33] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.
- [34] M. Santos, K. Santos, E. Alves, and A. Dantas, “CPU Bound Analysis of Wordcount Application in Hadoop Yarn Virtualized Nodes Using the Xen Platform,” in *2018 Symposium on High Performance Computing Systems (WSCAD)*, IEEE: pp. 274–274, 2019, doi:10.1109/wscad.2018.00056.
- [35] P.J. Morris, “The dawn of big data.,” in *North Carolina medical journal*, IEEE: pp. 177, 2014, doi:10.18043/ncm.75.3.177.S. Nishimura, S. Das, D. Agrawal, and A. El Abbadi, “MD-HBase: A scalable multi-dimensional data infrastructure for location aware services,” in *Proceedings - IEEE International Conference on Mobile Data Management*, IEEE: pp. 7–16, 2011, doi:10.1109/MDM.2011.41.
- [36] T. White, “Hadoop: The definitive guide 4th Edition,” *Online*, vol. 54, pp. 258, 2012, doi:citeulike-article-id:4882841.
- [37] T.T. Htay, and S. Phyu, “Improving the performance of Hadoop MapReduce Applications via Optimization of concurrent containers per Node,” in *2020 IEEE Conference on Computer Applications, ICCA 2020*, IEEE: pp. 1–5, 2020, doi:10.1109/ICCA49400.2020.9022836.
- [38] R.K. Chawda, and G. Thakur, “Big data and advanced analytics tools,” in *2016 symposium on colossal data analysis and networking (CDAN)*, IEEE: pp. 1–8, 2016.
- [39] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, and I. Stoica, “Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,” in *Proceedings of NSDI 2012: 9th USENIX Symposium on Networked Systems Design and Implementation*, pp. 15–28, 2012.
- [40] M. Armbrust, R.S. Xin, C. Lian, Y. Huai, D. Liu, J.K. Bradley, X. Meng, T. Kaftan, M.J. Franklinsky, A. Ghodsi, and M. Zaharia, “Spark SQL: Relational data processing in spark,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 1383–1394, 2015, doi:10.1145/2723372.2742797.
- [41] J. Kroß, and H. Krcmar, “Modeling and Simulating Apache Spark Streaming Applications,” *Softwaretechnik-Trends*, vol. 36, no. 4, pp. 1–3, 2016.
- [42] J.E. Gonzalez, R.S. Xin, A. Dave, D. Crankshaw, M.J. Franklin, and I. Stoica, “GraphX: Graph processing in a distributed dataflow framework,” in *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2014*, pp. 599–613, 2014.
- [43] Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D.B. Tsai, M. Amde, and S. Owen, “Millib: Machine learning in apache spark,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235–1241, 2016.
- [44] D. Lyubimov, and A. Palumbo, *Apache Mahout: Beyond MapReduce*, CreateSpace Independent Publishing Platform, 2016.
- [45] H. Schildt, “Big data and organizational design—the brave new world of algorithmic management and computer augmented transparency,” *Innovation: Management, Policy and Practice*, vol. 19, no. 1, pp. 23–30, 2017, doi:10.1080/14479338.2016.1252043.
- [46] P. Geczy, “Big data characteristics,” *The Macrotheme Review*, vol. 3, no. 6, pp. 94–104, 2014.

Security Aspects of Electronic Health Records and Possible Solutions

Prashant Vilas Kanade¹

Research Scholar, Department of Computer Science and Engineering, Sir Padampat Singhania University
Udaipur, Rajasthan, India

Dr. Arun Kumar²

Professor, Department of Computer Science and Engineering, Sir Padampat Singhania University
Udaipur, Rajasthan, India

Abstract—Health related information of a person in systematic format using information and Communication technology is definitely required. Storing patient information according to guidelines provided by government will help to achieve the concept of one person one record. There is also need to share the personal health record whenever necessary. If patient record (History) is readily available, it will help to make correct decisions related to patient's treatment. In our country (India) Ministry of Health and Family Welfare have recommended to eliminate conventional health record system. The major focus of this paper is to represent various methodologies that are adopted to implement web based health record system. As there is need of security while accessing and sharing of health related information, security is the major factor. Use of block chain, cryptography and timestamp based log record method is discussed. To assure the sharing of records, Inter Planetary File System (IPFS) is also discussed. Major purpose is to provide systematic and easy to use interoperable electronic health Record system.

Keywords—Patient history; cryptography; blockchain; timestamp based record; IPFS; electronic health records

I. INTRODUCTION

To provide an effective EHR System it is necessary to focus on some factors related to health care services. There are certain medical terms that definitely need to be studied. The various health record standards are also recommended by health authorities in India. Whenever there is a need to share patient's medical record to other medical expert or other health care centre it is necessary that it should be provided in standard format recommended by government health authority. If medical records are in unstructured or scattered format it is very difficult to analyze the record and that may extend the delay in decision making about the treatment. Web based system is easy to configure and easy to use system. This can be a conventional system to implement EHR system. Various Human Machine Interaction styles are adopted in web based system. The major disadvantage of the web based system is high demand of security arrangements as it has its own limitations. In India most of the medical practitioners are using fully specified names or locally identified terms while recording the diagnosis about the disease. Web based system is found feasible to record the healthcare information using standard terms and it will be feasible to share and analyze healthcare data [13]. Security related issues are definitely required to be addressed. There are various standards such as ICD, SNOMED-CT, LOINC, UML to represent patient's

disease information. HL7 and XML are one of the popular communication standards to share Patients information. These standards are beneficial to represent correct medical terms for patient data. Most of the doctors are using locally identified terms to represent patient health information. Very few practitioners have adopted ICD to represent the patient disease information. In India there is a rare scenario to maintain patient's health care information in standard format and in computer memory in electronic form. Web based system will also provide facility to share health record using communication standards such as XML and HL-7. Concept of hashing and block chain is emerging area in security of multiuser system [17].

II. INDIAN SCENARIO ABOUT HEALTH FACILITY

In India the health care system is decentralised. Health services hierarchy in India is from rural sector to urban sector. It is necessary to provide easy to use EHR system. EHR system should also assure meaningful use of data. In 2013, Ministry of Health and Family Welfare notified Electronic Health Records (EHR) Standards for India. The set of Standards given therein were selected from successful standards applicable to EHRs from around the world [14]. Detailed analysis is carried out about suitability and applicability of these standards in India by some expert group. Standards have been improvised and made according to the ever changing need of the mass. In these guidelines detailed recommendation on interoperability standards and clinical informatics standards, data ownership, privacy and security aspects are discussed.

Salient features [11] of electronic health record systems are i) Availability of Records ii) Summary of Clinical events iii) Evidenced Based Care. iv) Faster and Accurate diagnosis and v) Enhance the personalised care. There are certain challenges towards EHR in India. In the recent article from Hindustan Times it is mentioned that in India there are 1 Million doctors of Modern medicine to treat 1.3 Billion of its Population. There are hardly 1.5Lakh Doctors in Public service to serve patients. There is absolute non-existence of patient centric care in our Country. In India there are very less medical facility units to provide healthcare to huge population. Below mentioned table specifies the statistics of public health centres in India. We can observe from the below mentioned table 1 that, there is a growth in number of health centers in India, but these numbers are much below the requirement as far as population of India (133 Cr) is concerned.

TABLE I. STATISTICS OF HEALTH CENTRES IN INDIA

YEARS →	Health Centres in India				
	2005	2012	2017	2018	2019
Sub Centres	146026	148366	156231	158417	160713
Primary Health Centres (PHCs)	23236	24049	28863	25743	30045
Community Health Centres(CHCs)	3346	4833	5624	5624	5685

There are certain challenges such as lack of manpower, lack of infrastructure and lack of awareness among all service providers regarding proper recording of health care information [1].

In India there are various treatment methods such as allopathy, Homeopathy, Ayurveda and Unani. There are also various successful treatment methods which are found successful for some select diseases. Most of the people are dependent on general practitioners and traditional treatment methods in urban and semi-urban areas. It is necessary if all treatment related history to be recorded in standard format and that will definitely add a benefit to healthcare services in India. There is no any methodology adapted to record patient’s health care data.

Health records are generated at every health service centre. Most of the records are either lost or just lying in physical form with medical service unit or with the patient. Some records are destroyed after certain period [2].OPD record is normally handed over to patients. To make it varied purpose EHR is necessary in India. It can be made available for various users for various purposes. It can also be made available for all direct and indirect stakeholders. It is also necessary to provide Patient Centric health information system along with addressing Security and privacy issues. There is also need to address issues of ownership and governance.

III. DESIGN CONSIDERATIONS OF EHR SYSTEM

Following is a screen shot of primary model that comprise of web based recording of patient’s data. It is based on salient features recommended by ministry of health and family welfare (MOHFW) govt. of India. A Doctor is motivated from the EHR systems initiated by various countries worldwide.

Web Based Approach: To initiate the EHR system, a web Based system is recommended. In this initial approach PHP MY SQL and JASON is widely adopted. As shown in figure 1 below, we have provided facility to automatically insert ICD and SNOMED CT code for diseases. This is a system proposed for General Practitioners and same can be further stored in standard form on centralised accessible system.

Figure 2: Typical HTML Form to record Patient Case Information.

Doctors use fully Specified Name or Locally Identified name to record patient case. It is feasible for them to store record in less available time. In above web based system efforts are made to store the information in two standard formats for one FSN or Local Name of the disease. That

mapping will help to store disease information in ICD as well as in SNOMED code for each medicinal interaction. Mapping table is generated by including all necessary standards; one can definitely get benefit of using these standards. Following table is a compact version of the mapping table.

To create mapping of these standards special efforts need to be taken to locate the correct equivalent name and verify it be Subject Matter Experts [5]. It is mandatory to involve experts in this mapping process.

Pre-mapped disease names as shown in the table 2, will help to store the patient disease information with standard names (codes). Doctors can also retrieve the same in required format for particular Patient based on simple web based programs for the same [4]. Database and record will be generated. Health information can be accessed with SQL query. We can insert the key term in the text box and key specific data can be retrieved. Natural Language processing method of SQL and that will also help to access patient information based on some key terms. We can browse the patients with specific disease based on FSN, SNOMEDCT or UML key terms. Users can be provided with facility to type their query in simple English sentence as well as we can get their speech converted into text. Further that text can be used as an input string to access relevant data. In MySql NLP Full Text Search technique MySql will look for rows or documents that are relevant to the free-text natural human language query. Accessing Health data will be especially for researchers who wish to apply some analytics to the health data [6]. It will be also useful for organizations to take any strategic decision from the available systematic data. Figure 2 represents the interface to type a simple English question for particular information.

Fig. 1. User Interface to Record Patient Diagnosis.

TABLE II. MAPPING OF EHR STANDARDS

Fully Specified Name	ICD	UMLS	SNOMED CT
Cholera due to Vibrio cholerae 01, biovarcholerae	A000	C0494021	63650001
Cholera due to Vibrio cholerae 01, biovareltor	A001	C0343372	63650001
Cholera, unspecified	A009	C0008354	63650001

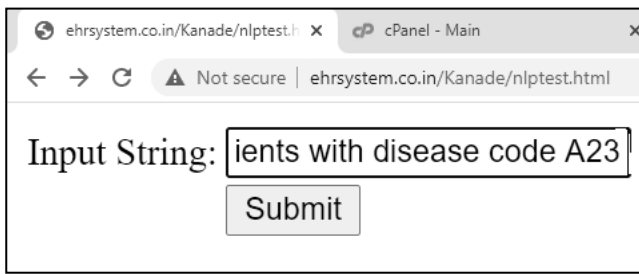


Fig. 2. NLP Search Technique (Input).

The basic goal of adopting NLP is to build intelligent system that will provide man-machine interface to understand speech and Text. Using NLP we can achieve computer aided Instruction Mechanism that can provide the information when and wherever required. It will also help to enhance the automatic storage and manage the information. NLP for EHR is mainly needed to extract healthcare information stored in unstructured form such as Files, Clinical Notes Reports etc. NLP search technique will search each keyword of the string and will provide the row of the matching words in the database will be retrieved [12]. This will also provide the specific record about the intended search criteria.

Figure 3 shows the output for the string that has a simple English question “I want Patients with disease code A23” NLP test technique will provide all the patients having disease code A23.

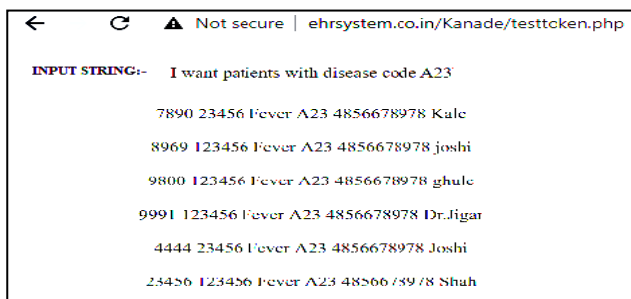


Fig. 3. NLP Search Technique (Output).

IV. SECURITY ISSUES

As far as security is concerned there are three major aspects of security as Human Factors, Technological Factors and Policy (Legal and Logical) factors. As it is a web based model there are certain limitations as far as security issues are faced while handling or accessing records. The major issues with respect to security are privacy, data breach and identity theft [3].

1) *Data breach*: There is a possibility that data may be shared unknowingly to the outside world through the devices such as smart phones or hand held devices such as tabs and palmtops.

There may be sharing of the data without any intention to malicious threats. If the information consist of any confidential information and that may be disclosed to unauthorized users. Some people will try to process the information either by updating or erasing important health related data. We may adopt some measure towards security and avoid the

unauthorised access but it is not possible to avoid the possibilities in which manner data will be accessed.

2) *Identity theft*: Once data is available in digital form there is a high risk of identity theft. Normally person will not be willing to share his health information to everyone who is on the digital network. Patient may share that data to his attending doctor, insurance provider and his family members. There is a possibility that data may be sold to similar service providers such as Private hospitals and insurance provider companies and pharmacy companies. Some data challengers may alter the historical data and then it will be very difficult to revert back to earlier version of information. It is observed that there are major challenges towards sharing of medical records. Following are the major challenges are with respect to interoperability. As we have mentioned earlier that Detailed and reliable workflows to share the data outside of the originating organization have not been established. Syntactic, Semantic and Process.

3) *Interoperability*: EHR is not just creation of the patient’s health information but it is necessary to Exchange of Information Between two Different Healthcare Systems. To provide a system that is model 1 with EHR Standards recommended by Government of India. Policy regarding Patient Identifiers and Coding Terms should be available [9].

Data can be stored in encrypted form as well as it can be made secured using the hashing and block chain technique [15]. Hashing is carried out and same is encapsulated using a block chain. To assure the protection of confidential health related data, Block chain method is used for each transaction that took place in patient’s health information [8]. When at each time when patient’s record is generated a unique key will be generated for each record if there is any update in health record the new key will be generated by encapsulating the previously generated unique key. Every time for each new transaction the block will be updated [16]. Same is represented in Figure 4.

4) *Time stamp recording of a data access*: To keep track of the interactions carried out with the patient record a timestamp based hashing is created [10]. Whenever patient record is accessed each time its hash key will be generated and stored separately in a database along with the timestamp. This is recommended just to keep track of changes (if any) in the record. So that any change in the record can be identified. Following table shows the typical timestamp based hashes for each accessed record. If there is a change in the record its hash value will change else there is no any change reflected in hash value. So any authorized or unauthorized access can also be identified for the health data.

As mentioned in the Table 3, one can observe that for patient id 532 there is access of record three times and when record was accessed on Dec 7, 2020 at 20:49 and at 20:50 at both this access points there is no change in hash; no change in record, but when this patient id was accessed and changed at 20:55 we can identify the change in hash. So from log record we can easily detect any change for the said record.

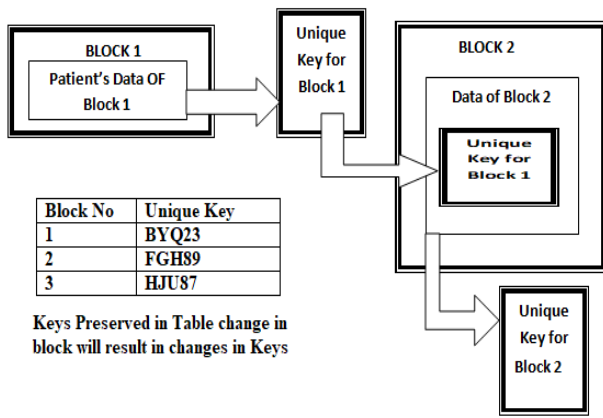


Fig. 4. Hashing to Assure Protected Access to Data using Block Chain.

TABLE III. TIME STAMP BASED HASH FOR PATIENT ID

Patient ID	HASH FUNCTIONS	TIME STAMP
532	da39a3ee5e6b4b0d3255bfef95601890afd80709	2020-12-07 20:49:57
532	da39a3ee5e6b4b0d3255bfef95601890afd80709	2020-12-07 20:50:32
896	33443a4d4e4336e039ef6ec31c55406214ef77b5	2020-12-07 20:53:01
896	33443a4d4e4336e039ef6ec31c55406214ef77b5	2020-12-07 20:53:39
896	33443a4d4e4336e039ef6ec31c55406214ef77b5	2020-12-07 20:54:19
532	306f02a37c290b4f7e8bb48ef2e1761e6a0a71ed	2020-12-07 20:54:52
896	33443a4d4e4336e039ef6ec31c55406214ef77b5	2020-12-07 20:55:19
532	306f02a37c290b4f7e8bb48ef2e1761e6a0a71ed	2020-12-07 20:55:42
1204	0b40883b8f1f3fb857054f6fcdcb35e82b5338942	2020-12-07 20:56:12
1204	0b40883b8f1f3fb857054f6fcdcb35e82b5338942	2020-12-07 20:58:10
896	33443a4d4e4336e039ef6ec31c55406214ef77b5	2020-12-07 21:01:10
532	306f02a37c290b4f7e8bb48ef2e1761e6a0a71ed	2020-12-07 21:03:02
5220	4dfa5899c4ebc3810090badcd194c87595c7110d	2020-12-10 20:12:41
5220	4dfa5899c4ebc3810090badcd194c87595c7110d	2020-12-10 20:14:00

V. INTERPLANETARY FILE SYSTEM

The Inter Planetary File System is a protocol and peer-to-peer network for storing and sharing data in a distributed file system. IPFS uses content-addressing to uniquely identify each file in a global namespace connecting all computing devices. IPFS can be used to share the patient specific information. The major advantage of IPFS is that it creates an IPFS key, normally called as IPFS key or also called as cryptographic Hash. In health records if we want to share images, health reports or a specific patient information, then we can use ipfs to

create its compress key and send this key to the intended user. The recipient can use IPFS to retrieve the file in its original form. In this research we have tested IPFS for sharing DICOM (file with extension .dcm) images. A file having size of 1.62 MB is converted into a file as an encrypted string with a smaller size of only 1 KB or less than that. IPFS has represented the huge file knees.dcm to a small string as “QmcnteXR2CLXy6bSMU97iMum9ppmMriq7kaFvs8qCmA.JAL” this string can be send to the recipient and at the recipient end this can be regenerated using ipfs I/O environment. Figure 5 represents the detailed description of IPFS to share the files.

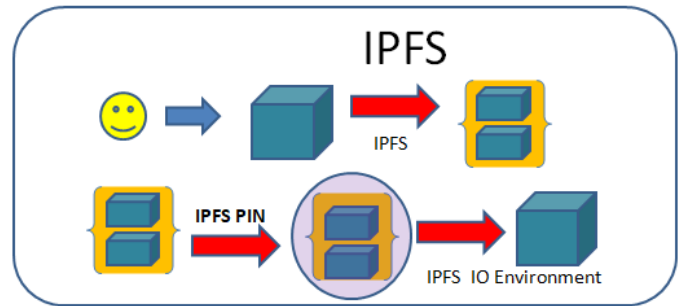


Fig. 5. Interplanetary File System (IPFS) to Share the Data.

Only problem is that if there is a need to protect the file from public viewing this IPFS has some limitations. We can encrypt the file and then share to keep the confidentiality of the content. Figure 6 gives a stepwise representation of encrypting the file. There are two approaches as either encrypt the file or encrypt the string generated by the IPFS. As shown in the figure a file is encrypted first and then its encrypted form will be converted into IPFS key. It will be forwarded to the recipient to assure the privacy of the file.

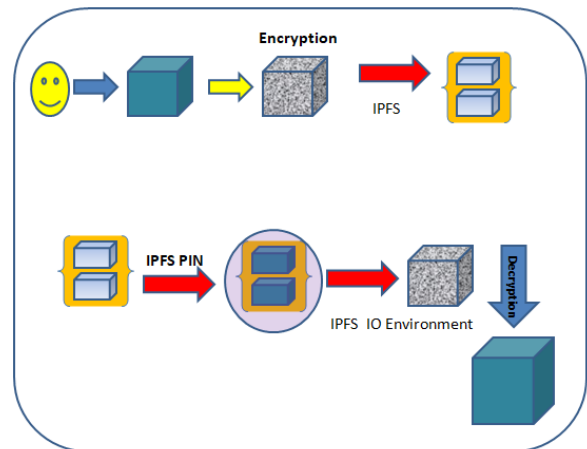


Fig. 6. IPFS with Encryption of Data towards Confidentiality.

VI. CONCLUSION AND FUTURE WORK

There is a definite need to use Various Electronic Health Record Standards to create easy to use Web Based EHR System. A knowledge base of the EHR standards can be implemented where mapping of disease names can be made available so that it will help doctors to generate systematic Health records. It is need of the time to insist on Patient’s

Healthcare Information along with EHR standards [7]. To promote globally sharing of health records communication Standards like HL7 and XML are suitable for the same. We can easily access the Patient's Information. To make it more feasible NLP Full Text method is found suitable to access specific data of a patient.

A major challenge is there to providing a role based access to avoid data challenges. Real Time and Instant Recording of Data using smart devices is future scope of this research. It is also recommended to make it suitable for some Ayurvedic Terms. To motivate the intended users it is recommended to Train the users of the system.

REFERENCES

- [1] Bhartiya S., Mehrotra D., 2016, Girdhar A. Issues in Achieving Complete Interoperability while Sharing Electronic Health Records , International Conference on Information Security & Privacy (ICISP2015), Nagpur, INDIA , (2016) Procedia Computer Science, 78 , pp. 192-198.
- [2] Bhattacharya I and Muju S (2012), Need for Interoperability standards for Healthcare in India, CSI Journal of Computing, VOL 1. No.4.
- [3] Calvanese D, Giacomo G De, Domenico L, Data Management in Peer to peer Data Integration Systems uploaded on https://www.researchgate.net/profile/Diego_Calvanese.
- [4] David W. The Road to Implementation of Electronic Health Record, Baylor University Medical Center Proceedings 19:4,311-312, DOI: 10.1080/08998280.2006.11928189.
- [5] EHR Standards for India(1), 2016, Guidelines given by Ministry of Health and Family welfare, e-governance Division.
- [6] EHR Standards for India(2), standards at a glance on a portal of National Resource Centre for EHR Standards shared on https://www.nrce.in/standards-for-India#standards_at_a_glance page 1-8.
- [7] Jian W, Wen H, Scholl J., Shabbir S, Lee P., Hsu C, Li Y, 2011, The Taiwanese method for providing patients data from multiple hospital EHR systems, Journal of Biomedical Informatics, 44 (2) , pp. 326-332.
- [8] Kuo T, Kim H, Ohno-Machado L. 2017, Blockchain distributed ledger technologies for biomedical and health care applications. J Am Med Inform Assoc;24(6):1211-1220. [doi: 10.1093/jamia/ocx068] [Medline: 29016974].
- [9] Mantri M, Gaur S, Sinha P (2010), "Model and Process Interoperability between Clinical Standards", Conference: 6th International Conference of Telemedicine Society of India (Telemedicon'10) At: Bhubaneswar, Odisha, <https://doi.org/10.13140/2.1.1390.8808>.
- [10] Meinert E, Alturkistani A, Foley KA, Osama T, Car J, Majeed A, Van Velthoven M, Wells G, Brindley D, 2019 Blockchain Implementation in Health Care: Protocol for a Systematic Review, JMIR Res Protoc ;8(2):e10994 , URL: <https://www.researchprotocols.org/2019/2/e10994> DOI: 10.2196/10994, PMID: 30735146, PMCID: 6384534.
- [11] Raza M, Good Medical 2012, Record Keeping, *International Journal of Collaborative Research on Internal Medicine & Public Health*, Volume 4, Issue 5, Pages 535-543. <http://internalmedicine.imedpub.com/>.
- [12] Sadvoski A, 2017, Simple NLP Search in your Application-step by step guide, <http://www.tech.evojam.com>.
- [13] Sinha P, Gaur S, Bendale P, Mantri M, Dande A (2012) Book Title Electronic Health Record, Standards, Coding Systems, Frameworks, and Infrastructures, IEEE Press ISBN 978-1-118-28134-5.
- [14] Srivastava, S. (2016). "Adoption of Electronic Health Records: A roadmap for India", Healthcare Informatics Research, 22(4), 261-269 <https://doi.org/10.4258/hir.2016.22.4.261>.
- [15] Tsung-Ting Kuo, Hyeon-Eui Kim, Lucila Ohno-Machado, Blockchain distributed ledger technologies for biomedical and health care applications, *Journal of the American Medical Informatics Association*, Volume 24, Issue 6, November 2017, Pages 1211-1220, <https://doi.org/10.1093/jamia/ocx068>.
- [16] Tejpal K, 2019 Blockchain: An Answer to Public EHR System, a Blog on <https://www.investindia.gov.in/team-india-blogs/blockchain-answer-public-electronic-health-record-ehr-system>.
- [17] Jim Atherton, MD, Virtual Mentor, 2011, HISTORY OF MEDICINE:186-189. doi: 10.1001/virtualmentor.2011.13.3.mhst1-1103, American Medical Association Journal of Ethics, Volume 13, Number 3: 186-189.